140.651 Problem Set 3 Solutions

Homework 2, Problem 10

When at the free-throw line for two shots, a basketball player makes at least one free throw 90% of the time. 80% of the time, the player makes the first shot, while 70% of the time she makes both shots.

a. Does it appear that the player's second shot success is independent of the first?

Define the events $S_1 = \{\text{First shot made}\}$ and $S_2 = \{\text{Second shot made}\}$. We are given $P(S_1 \cup S_2) = 0.9$, $P(S_1) = 0.8$, $P(S_1 \cap S_2) = 0.7$. By the law of total probability, $P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)$ implying that,

$$P(S_2) = P(S_1 \cup S_2) - P(S_1) + P(S_1 \cap S_2) = 0.9 - 0.8 + 0.7 = 0.8$$

For S_1 and S_2 to be independent, $P(S_1 \cap S_2) = P(S_1) \cdot P(S_2)$. However, $P(S_1) \cdot P(S_2) = 0.8 \cdot 0.8 = 0.64$ and $0.7 \neq 0.64$, so S_1 and S_2 are **not** independent.

b. What is the conditional probability that the player makes the second shot given that she made the first? What would it be if she missed the first?

The conditional probability that the player makes the second shot given that she made the first is $P(S_2|S_1)$. By definition of conditional probability,

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{0.7}{0.8} = 0.875$$

The conditional probability that the player makes the second shot given that she missed the first is $P(S_2|S_1^c)$. By the definition of conditional probability and complement, we have

$$P(S_2|S_1^c) = \frac{P(S_2 \cap S_1^c)}{P(S_1^c)} = \frac{P(S_2) - P(S_2 \cap S_1)}{1 - P(S_1)} = \frac{0.8 - 0.7}{1 - 0.8} = \frac{0.1}{0.2} = 0.5$$

Homework 2, Problem 15

A web site (www.medicine.ox.ac.uk/bandolier/band64/b64-7.html) for home pregnancy tests cites the following:

> When the subjects using the test were women who collected and tested their own samples, the overall sensitivity was 75%. Specificity was also low, in the range 52% to 75%.

a. Interpret a positive and negative test result using diagnostic likelihood ratios using both extremes of the specificity.

When the specificity $= 0.52$,

- The diagnostic likelihood ratio for a positive test result (DLR_+) is given by,

$$DLR_+ = \frac{P(+|D)}{P(+|D^c)} = \frac{0.75}{1 - 0.52} = 1.5625$$

  Thus, we can interpret the postive test result as: the post-test odds for pregnancy has 1.56x more support than the pre-test odds.

- The diagnostic likelihood ratio for a negative test result (DLR_-) is given by,

$$DLR_- = \frac{P(-|D)}{P(-|D^c)} = \frac{0.25}{0.52} = 0.4808$$

Thus, we can interpret the negtive test result as: the post-test odds for pregnancy has 0.48x the support of the pre-test odds.

When the specificity $= 0.75$,

- The diagnostic likelihood ratio for a positive test result (DLR_+) is given by,

$$DLR_+ = \frac{P(+|D)}{P(+|D^c)} = \frac{0.75}{1 - 0.75} = 3$$

Thus, we can interpret the postive test result as: the post-test odds for pregnancy has 3x more support than the pre-test odds.

- The diagnostic likelihood ratio for a negative test result (DLR_-) is given by,

$$DLR_- = \frac{P(-|D)}{P(-|D^c)} = \frac{0.25}{0.75} = 0.3333$$

Thus, we can interpret the negtive test result as: the post-test odds for pregnancy has 0.33x the support of the pre-test odds.

b. A woman taking a home pregnancy test has a positive test. Draw a graph of the positive predictive value by the prior probability (prevalence) that the woman is pregnant. Assume the specificity is 63.5%

To find the PPV assuming the specificity is 64.5%, we can use the formula,

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

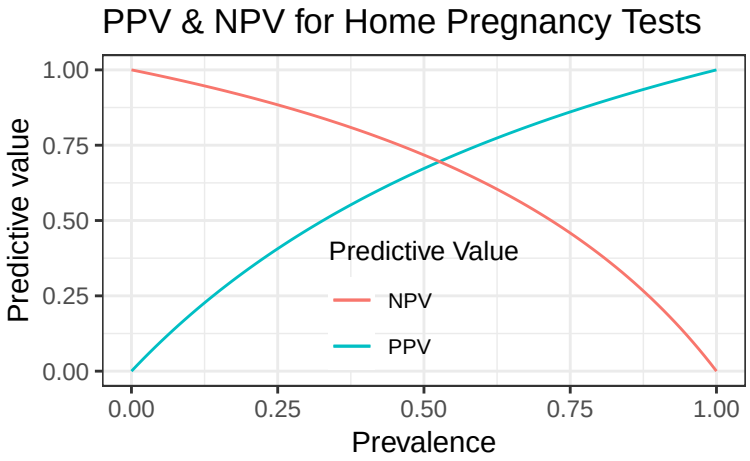
We will plot the PPV along with the NPV below in part (c).

c. Repeat the previous question for a negative test and the negative predictive value.

To find the negative predictive value (NPV), we can use the formula,

$$P(D^c|-) = \frac{P(-|D^c)P(D^c)}{P(-|D^c)P(D^c) + P(-|D)P(D)}$$

```
prev <- seq(0,1, len = 100)    # Prevalence
sens <- 0.75                   # Sensitivity
spec <- 0.635                  # Specificity
PPV <- sens * prev/(sens*prev + (1-spec)*(1-prev))          # PPV
NPV <- spec * (1- prev)/(spec*(1-prev) + (1 - sens)* prev)  # NPV

ggplot(data = data.frame(Prevalence = prev, PPV, NPV)) +
  geom_line(aes(x = prev, y = PPV, color = 'PPV')) +
  geom_line(aes(x = prev, y = NPV, color = 'NPV')) +
  labs(title = "PPV & NPV for Home Pregnancy Tests",
       x = "Prevalence", y = "Predictive value", color = "Predictive Value") +
  theme_bw() +
  theme(legend.position = c(0.5, 0.25),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        legend.background = element_rect(fill = alpha('white', 0 )))
```



Problem 1

Imagine that a person, say his name is Flip, has an oddly deformed coin and tries the following experiment. Flip flips his coin 10 times, 7 of which are heads. You think maybe Flip's coin is biased towards having a greater probability of yielding a head than 50%.

a. What is the maximum likelihood estimate of p, the true probability of heads associated with this coin?

Let X be a random variable for the number of heads flipped. Then, X follows a Binomial distribution with parameter p. Additionally, let x denote the observed outcome of flipping 7 heads. Finding the MLE for p,

$$L(p|x) = \binom{10}{x} p^x (1-p)^{10-x}$$

$$\log L(p|x) = \log \binom{10}{x} + x \log(p) + (10-x)\log(1-p)$$

$$\frac{d}{dp} \log L(p|x) = \frac{x}{p} - \frac{10-x}{1-p}$$

Setting the derivative of the log-likelihood equal to zero and solving for p, we get

$$\hat{p} = \frac{x}{10} = \frac{7}{10}$$

To verify that $\hat{p}$ maximizes the likelihood function, we can look at the sign of the second derivative of the log-likelihood at $\hat{p}$. The second derivative is given by,

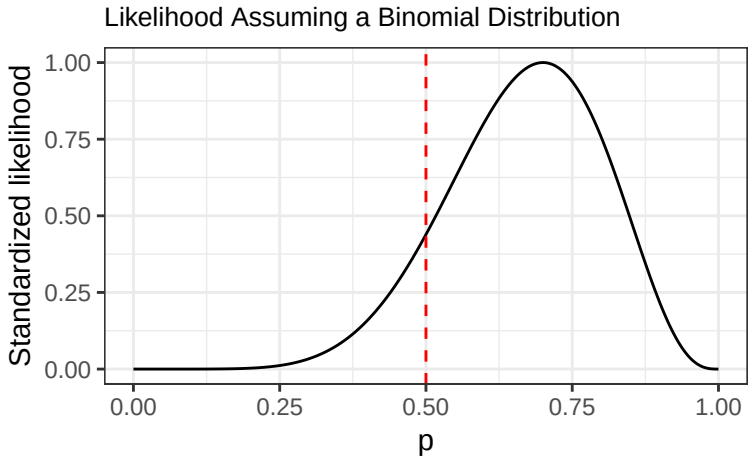
$$\frac{d^2}{dp^2} \log L(p|x) = -\frac{x}{p^2} - \frac{10-x}{(1-p)^2}$$

Evaluating at $p = \hat{p}$ and $x = 7$, we see that the second derivative of the log-likelihood is -47.62 which is negative. Thus, $\hat{p}$ is indeed the MLE for p.

b. Plot the likelihood associated with this experiment. Renormalize the likelihood so that its maximum is one. Does the likelihood suggest that the coin is fair?

```
p <- seq(0, 1, by=0.001)                    # Vector of probabilities
likelihood <-choose(10,7)*p^7 * (1 - p)^3   # Likelihood for each probability
stdlike<-likelihood/max(likelihood)         # Renormalize likelihood
ggplot(data = data.frame(p, stdlike), aes(x = p, y = stdlike)) +
  geom_line() +
  geom_vline(xintercept = 0.5, colour = "red", linetype = "dashed")+
```

```
  labs(title ="Likelihood Assuming a Binomial Distribution",
     x = "p", y = "Standardized likelihood") +
  theme_bw() +
  theme(plot.title = element_text(size = 10))
```



Given the data, the normalized likelihood that p is 0.5 is around 0.4. While the maximum likelihood estimate of 0.7 suggests that the coin is not fair, 0.5 appears to be within the range of plausible values of p.

c. Whats the probability of seeing 7 or more heads out of ten coin flips if the coin was fair? Does this probability suggest that the coin is fair? Note this number is called a P-value.

Assuming the coin was fair, we have $X \sim Binomial(n = 10, p = 0.5)$. Then, the probability of seeing 7 or more heads out of 10 coin flips if the coin was fair is given by,

$$P(X \geq 7) = \sum_{x=7}^{10} P(X = x)$$

We can evaluate this probability using the `pbinom()` function in `R`. The default argument in `pbinom()` is `lower.tail = TRUE` which calculates $P(X \leq x)$. However, setting `lower.tail = FALSE` gives us $P(X > x)$.

```
# Using pbinom with with the default lower.tail = TRUE
1 - pbinom(6, 10, .5 )
```

```
[1] 0.171875
```

```
# Using pbinom with lower.tail = FALSE
pbinom(6, 10, .5, lower.tail=FALSE)
```

```
[1] 0.171875
```

Assuing the coin is fair, we would see seven or more heads in 10 flips 17% of the time. Equivalently, the probability that we see an event as rare as flipping heads seven times is 17%. Using a threshold of 5%, there is not enough evidence to suggest that the coin is not fair.

d. Suppose that Flip told you that he did not fix the number of trials at 10. Instead, he told you that he had flipped the coin until he obtained 3 tails and it happened to take 10 trials to do so. Therefore, the number 10 was random while the number three 3 fixed. The probability mass function for the number of trials, say y, to obtain 3 tails (called the negative binomial distribution) is

$$\binom{y-1}{2}(1-p)^3 p^{y-3}$$

for $y = 3, 4, 5, 6, \ldots$. What is the maximum likelihood estimate of p now that we've changed the underlying mass function?

Let Y be a random variable denoting the number of flips until we see 3 tails, where Y follows a negative binomial distribution. Here, y is our observed outcome of 10 flips. Deriving the MLE of p given y,

$$L(p|y) = \binom{y-1}{2}(1-p)^3 p^{y-3}$$

$$\log L(p|y) = \log \binom{y-1}{2} + 3\log(1-p) + (y-3)\log(p)$$

$$\frac{d}{dp}\log L(p|y) = -\frac{3}{1-p} + \frac{y-3}{p}$$

Setting the derivative of the log-likelihood equal to 0 and solving for p, we have

$$\hat{p} = 1 - \frac{3}{y} = 1 - \frac{3}{10} = 0.7$$

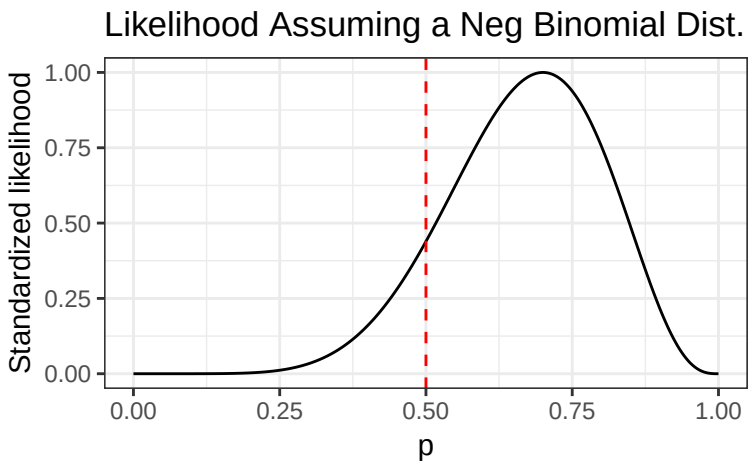
To verify that $\hat{p}$ maximizes the likelihood function, we can look at the sign of the second derivative of the log-likelihood at $\hat{p}$. The second derivative is given by,

$$\frac{d^2}{dp^2}\log L(p|y) = -\frac{3}{(1-p)^2} - \frac{y-3}{p^2}$$

Evaluating at $p = \hat{p}$, the second derivative or the log-likelihood is -47.62 which is negative, so $\hat{p}$ is indeed the MLE for p.

e. Plot the likelihood under this new mass function. Renormalize the likelihood so that its maximum is one. Does the likelihood suggest that the coin is fair?

```
p <- seq(0, 1, by=0.001)                          # Vector of probabilities
likelihood <-choose(10-1,2)*(1-p)^3 * p^(10-3)    # Likelihood for each probability
stdlike<-likelihood/max(likelihood)               # Renormalize likelihood

ggplot(data = data.frame(p, stdlike), aes(x = p, y = stdlike)) +
  geom_line() +
  geom_vline(xintercept = 0.5, colour = "red", linetype = "dashed")+
  labs(title ="Likelihood Assuming a Neg Binomial Dist.",
     x = "p", y = "Standardized likelihood") +
  theme(plot.title = element_text(size = 12)) +
  theme_bw()
```



Note that the likelihood curve is the same as in 1b, implying that the evidence about the coin being fair has not changed – despite the MLE estimate of $\hat{p} = 0.7$, 0.5 is still within the reasonable range of values.

f. Calculate the probability of requiring 10 or more flips to obtain 3 tails if the coin was fair. (Notice that this is the same as the probability of obtaining 7 or more heads to obtain 3 tails.) This is the P-value under the new mass function.

(Aside) This problem highlights a distinction between the likelihood and the P-value. The likelihood and the MLE are the same regardless of the experiment. That is to say, the likelihood only seems to care that you saw 10 coin flips, 7 of which were heads. Flip's intention about when he stopped flipping the coin, either at 10 fixed trials or until he obtained 3 tails, are irrelevant as far as the likelihood is concerned. The P-value, in comparison, does depend on Flip's intentions

Assuming the coin was fair, we have Y follows a negative binomial distribution with probability of success of 0.5 and number of failures as 3. Then, the probability of flipping the coin ten times before seeing 3 failures is given by,

$$P(Y \geq 10) = 1 - P(Y < 10) = 1 - \sum_{y=1}^{9} P(Y = y)$$

We can evaluate this probability using the `pnbinom()`. Note `R` has a different parametrization for the negative binomial where the random variable is the number of success before failures. Again, we can use the argument `lower.tail = FALSE` to get $P(Y \geq 10) = P(Y > 9)$.

```
# Under R's parametrization for the negative binomial,
#     The first parameter is the number of successes before failures: 9 - 3 = 6
#     The second parameter is the number of failures: 3
pnbinom(6, 3, .5, lower.tail=FALSE)
```

```
[1] 0.08984375
```

Notice that the p-value found under the negative binomial distribution is different from the p-value found under the binomial distribution.

Problem 2

Suppose a researcher is studying the number of sexual acts with an infected person until an uninfected person contracts an sexually transmitted disease. She assumes that each encounter is an independent Bernoulli trial with probability p that the subject becomes infected. This leads to the so-called geometric distribution $P(\text{Person is infected on contact } x) = p(1-p)^{x-1}$ for $x = 1, \ldots$.

a. Suppose that one subject's number of encounters until infection is recorded, say x. Symbolically derive the ML estimate of p.

Deriving the MLE for p,

$$\begin{aligned} L(p|x) &= p(1-p)^{x-1} \\ \log L(p|x) &= \log(p) + (x-1)\log(1-p) \\ \frac{d}{dp}\log L(p|x) &= \frac{1}{p} - \frac{x-1}{1-p} \end{aligned}$$

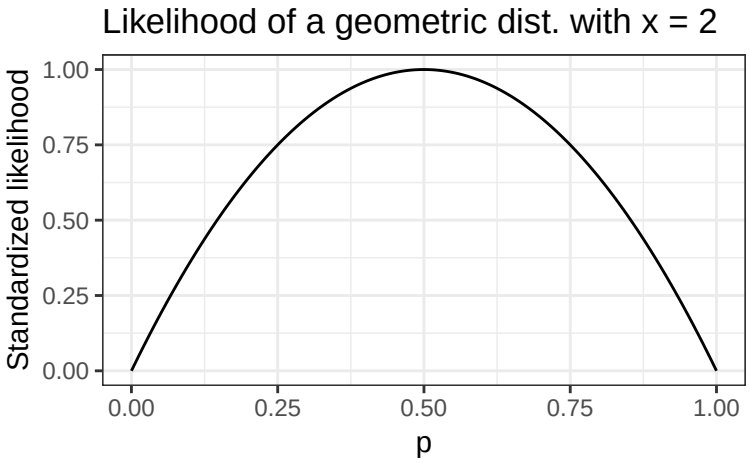
Setting the derivative of the log-likelihood equal to zero and solving for p, we get $\hat{p} = \frac{1}{x}$. To verify that $\hat{p}$ maximizes the likelihood function, we can look at the sign of the second derivative of the log-likelihood at $\hat{p}$. The second derivative is given by,

$$\frac{d^2}{dp^2}\log L(p|x) = -\frac{1}{p^2} - \frac{x-1}{(1-p)^2}$$

The second derivative is negative as $-\frac{1}{p^2}$ is always negative and $-\frac{x-1}{(1-p)^2} \leq 0$ since $x \geq 1$. Thus, $\hat{p}$ is indeed the MLE for p.

b. Suppose that the subjects value was 2. Plot and interpret the likelihood for p.

```
p <- seq(0, 1, by=0.001)                # Vector of probabilities
likelihood <-p*(1-p)^(2-1)              # Likelihood for each probability
stdlike<-likelihood/max(likelihood)     # Renormalize likelihood

ggplot(data = data.frame(p, stdlike), aes(x = p, y = stdlike)) +
  geom_line() +
  labs(title ="Likelihood of a geometric dist. with x = 2",
     x = "p", y = "Standardized likelihood") +
  theme(plot.title = element_text(size = 6)) +
  theme_bw()
```



The maximum likelihood estimate for p is $\hat{p} = \frac{1}{2}$. That is, given we observed two encounters until infection, the estimate for the true probability of infection per encounter that maximizes the likelihood is 0.5.

c. Suppose that is often assumed that the probability of transmission, p, is .01. The researcher thinks that it is perhaps strange to have a subject get infected after only 2 encounters if the probability of transmission is really only 1%. According to the geometric mass function, what is the probability of a person getting infected in 2 or fewer encounters if p truly is .01?

Let $X \sim Geometric(p)$ where X is the number of encounters until infection. Then, the probability of a person getting infected in 2 or fewer encounters is,

$$\begin{aligned} P(X \leq 2) &= P(X = 1) + P(X = 2) \\ &= (0.01) * (1 - 0.01)^{1-1} + (0.01) * (1 - 0.01)^{2-1} \\ &= 0.01 + 0.0099 = 0.0199 \end{aligned}$$

Checking our answer in `R` using the `pgeom()` function,

```
# pgeom() parametrizes with number of failures before success
#   We are interested in infection after at most 2 encounters
#   which means infection after at most 1 failure.
pgeom(1, prob = 0.01)
```

```
[1] 0.0199
```

d. Suppose that she follows n subjects and records the number of sexual encounters until infection (assume all subjects became infected) $x_1, \cdots, x_n$. Symbolically derive the ML estimate of p.

Deriving the MLE for p,

$$\begin{aligned} L(p|x_1, \cdots, x_n) &= \prod_{i=1}^{n} L(p|x_i) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum_i (x_i - 1)} \\ &= p^n (1-p)^{\sum_i x_i - n} \\ \log L(p|x_1, \cdots, x_n) &= n \log(p) + \left(\sum_{i=1}^{n} x_i - n \right) \log(1-p) \\ \frac{d}{dp} \log L(p|x_1, \cdots, x_n) &= \frac{n}{p} - \frac{\sum_i x_i - n}{1-p} \end{aligned}$$

Setting the derivative of the log-likelihood equal to 0 and solving for p, we get that the maximum likelihood estimate of p is given by,

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{n/n}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

To verify that $\hat{p}$ maximizes the likelihood function, we can look at the sign of the second derivative of the log-likelihood at $\hat{p}$. The second derivative is given by,

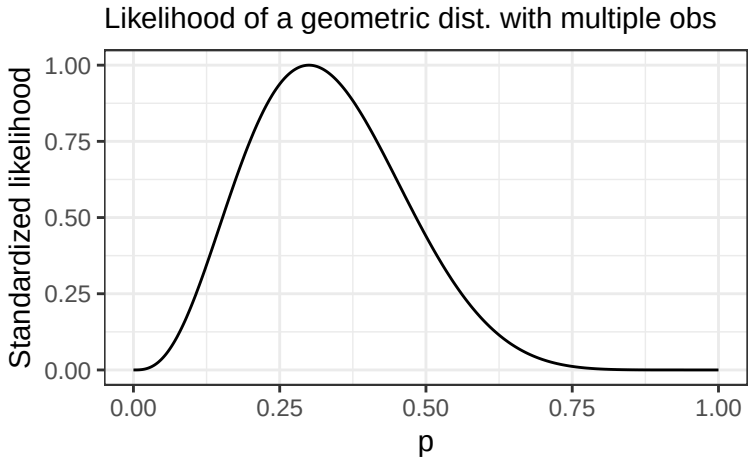
$$\frac{d^2}{dp^2} \log L(p|x_1, \cdots, x_n) = -\frac{n}{p^2} - \frac{\sum_{i=1}^{n} x_i - n}{(1-p)^2}$$

Since $n \geq 1$ and $\sum_{i=1}^{n} x_i \geq n$ since $x_i \geq 1$ for all i, we have $-\frac{n}{p^2} \leq 0$ and $-\frac{\sum_i x_i - n}{(1-p)^2} \leq 0$, implying that the second derivative is negative. Thus, $\hat{p}$ is indeed the MLE for p.

e. Suppose that she records values $x_1 = 3$, $x_2 = 5$, $x_3 = 2$. Plot and interpret the likelihood for p.

Here, $n = 3$ and $\sum_{i=1}^{3} x_i = 3 + 5 + 2 = 10$.

```
p <- seq(0, 1, by=0.001)                  # Vector of probabilities
likelihood <-p^(3)*(1-p)^(10-3)           # Likelihood for each probability
stdlike<-likelihood/max(likelihood)       # Renormalize likelihood
ggplot(data = data.frame(p, stdlike), aes(x = p, y = stdlike)) +
  geom_line() + labs(title ="Likelihood of a geometric dist. with multiple obs",
     x = "p", y = "Standardized likelihood") +
  theme_bw() + theme(plot.title = element_text(size = 11))
```



The maximum likelihood estimate for p is $\hat{p} = 0.3$. That is, given that we observed 3, 5, and 2 encounters until infection, the most likely estimate for the true probability of infection per encounter is 0.3.

Problem 3

In a study of aquaporins 6 frog eggs received a protein treatment. If the treatment of the protein is effective, the frog eggs would implode. The experiment resulted in 5 frog eggs imploding. Historically, ten percent of eggs implode without the treatment. Assuming that the results for each egg are independent and identically distributed:

a. What's the probability of getting 5 or more eggs imploding in this experiment if the true probability of implosion is 10%? Interpret this number.

Let X denote the number of eggs imploding in this experiment and assume $X \sim Binomial(6, p)$. Assuming $p = 0.5$, we have

$$P(X \geq 5) = P(X = 5) + P(X = 6) = \binom{6}{5} 0.1^5 (1 - 0.1)^{6-5} + \binom{6}{6} 0.1^6 (1 - 0.1)^{6-6} = 5.5 \times 10^{-5}$$

Verifying our answer using the function `pbinom()` in `R`,

```
# Recall lower.tail = FALSE gives P(X > x)
pbinom(4, 6, 0.1, lower.tail = FALSE)
```

```
[1] 5.5e-05
```

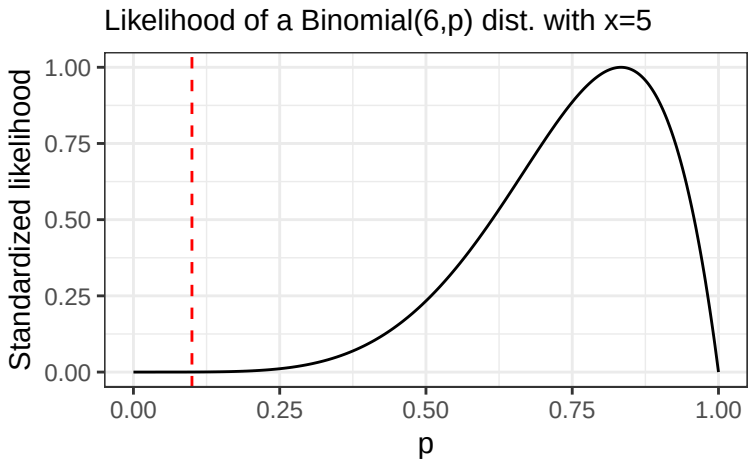
b. What is the maximum likelihood estimate for the probability of implosion?

In problem 1, we showed that, given one observation for a binomial random variable with parameters n, p, the maximum likelihood estimate of p is given by,

$$\hat{p} = \frac{x}{n} = \frac{5}{6}$$

c. Plot and interpret the likelihood for the probability of implosion.

```
p <- seq(0, 1, by=0.001)                  # Vector of probabilities
likelihood <-choose(6,5)*p^5*(1-p)        # Likelihood for each probability
stdlike<-likelihood/max(likelihood)       # Renormalize likelihood
ggplot(data = data.frame(p, stdlike), aes(x = p, y = stdlike)) +
  geom_line() + geom_vline(xintercept = 0.1, colour = "red", linetype = "dashed") +
  labs(title ="Likelihood of a Binomial(6,p) dist. with x=5",
     x = "p", y = "Standardized likelihood") +
  theme_bw() + theme(plot.title = element_text(size = 11))
```



The maximum likelihood estimate for p is $\hat{p} = \frac{5}{6}$. That is, given that we observed five eggs of the six eggs exploding, the maximum likelihood estimate for the true probability of infection per encounter is $\frac{5}{6}$. In the plot of the likelihood we see that the likelihood of the true probability of eggs exploding being 0.1 is very low.

Problem 4

(Adapted from Rosner page 135) Suppose that the diastolic blood pressures of $35-44$ year old men are normally distributed with mean 80 ($mm\ Hg$) and variance 144. For the same population, the systolic blood pressures are also normally distributed and have a mean of 120 and variance 121.

a. What is the probability that a randomly selected person from this population has a DBP less than 90?

For $35-44$ year old men, let X be the diastolic blood pressure where $X \sim N(\mu_X = 80, \sigma_X^2 = 144)$, and Y be the systolic blood pressure where $Y \sim N(\mu_Y = 120, \sigma_Y^2 = 121)$. Then, the probability that a randomly selected person has a DBP less than 90 is given by, $P(X \leq 90)$. We can use the `pnorm()` function in `R` to find this probability.

```
pnorm(90, mean = 80, sd = 12)
```

```
[1] 0.7976716
```

b. What DBP represents the 90^{th}, 95^{th} and 97.5^{th} percentiles of this distribution?

Using the `qnorm()` function in `R`, the 90^{th}, 95^{th} and 97.5^{th} percentiles are given by,

```
percentile <- qnorm(c(.90, .95, .975), mean=80, sd=12)
```

c. What's the probability of a random person from this population having a SBP 1, 2 or 3 standard deviations above 120? What's the corresponding probabilities for having DBPs 1, 2 or 3 standard deviations above 80?

Using the `pnorm()` function in `R`,

```
# SBP
mean.s  <- 120
var.s   <- 121
sd.above <- c(mean.s + sqrt(var.s), mean.s + 2*sqrt(var.s), mean.s + 3*sqrt(var.s))
round(pnorm(sd.above, mean.s, sqrt(var.s), lower.tail=FALSE), 4)
```

```
[1] 0.1587 0.0228 0.0013
```

```
# DBP
mean.d  <- 80
var.d   <- 144
sd.above <- c(mean.d + sqrt(var.d), mean.d + 2*sqrt(var.d), mean.d + 3*sqrt(var.d))
round(pnorm(sd.above, mean.d, sqrt(var.d), lower.tail=FALSE), 4)
```

```
[1] 0.1587 0.0228 0.0013
```

Thus, we see that for both DBP and SBP,

$$P(1 \text{ sd above mean}) = 15.87\%$$
$$P(2 \text{ sd above mean}) = 2.28\%$$
$$P(3 \text{ sd above mean}) = 0.13\%$$

d. Suppose that 10 people are sampled from this population. What's the probability that 50% (5) of them have a SBP larger than 140? –>

Let $Y_1, Y_2, \cdots, Y_{10}$ be the SBP of the 10 people sampled from the population. Furthermore, for $i = 1, 2, \cdots, 10$, define,

$$Z_i \begin{cases} 1 & \text{if } Y_i > 140 \\ 0 & \text{if } Y_i \leq 140 \end{cases}$$

That is, Z_i is an indicator variable for whether individual i has SBP larger than 140. Then, we are interested in $P\left(\sum_{i=1}^{10} Z_i = 5\right)$. Since we assume that the individuals are sampled independently, we know that $\sum_{i=1}^{10} Z_i \sim Binomial(10, p)$ where $p = P(Y_i > 140)$.

```
# Find P(Y_i > 140)
p <- pnorm(140, 120, 11, lower.tail = FALSE)

# Find P(sum Z_i = 5)
dbinom(5, 10, prob=p)
```

```
[1] 1.036001e-05
```

e. Suppose that 1,000 people are sampled from this population. What's the probability that 50% (500) of them have a SBP larger than 140?

We can use the same approach as in part (d). Now, let $Y_1, Y_2, \cdots, Y_{1000}$ denote the SBP of the 1,000 sampled individuals, and $Z_1, Z_2, \cdots, Z_{1000}$ denote the indicator variable as defined in part (d). Now, we are interested in $P\left(\sum_{i=1}^{1000} Z_i = 5000\right)$. Again, since we assume that the individuals are sampled independently, we know that $\sum_{i=1}^{1000} Z_i \sim Binomial(1000, p)$ where $p = P(Y_i > 140)$.

$P(Y_i > 140)$ is the same as in part (d), so using the `dbinom()` function, we get that the probability that 50% of the 1,000 sampled people have SBP larger than 140 is given by,

```
dbinom(500,1000, prob=p)
```

```
[1] 0
```

While `dbinom()` returns a value approximately 0, the true probablity is not actually 0, but extremeley close to 0.

f. If a person's SBP and DBP are independent, what's the probability that a person has a SBP larger than 140 and a DBP greater than 90? Is independence a good assumption?

Assuming that SBP and DBP are independent, we get that

$$P(X_i > 90, Y_i > 140) = P(X_i > 90)P(Y_i > 140)$$

Using the `pnorm()` function in `R` to calculate the probability, we get,

```
pnorm(90,80,12, lower.tail=FALSE)*pnorm(140,120,11, lower.tail=FALSE)
```

```
[1] 0.006984006
```

Independence is not a good assumption here because SBP and DBP change in the same direction for each individual. As SBP increases, we expect DBP to also increase, and vice versa.

g. Suppose that an average of 200 people are drawn from this population. What's the probability that this average is smaller than 81.3?

Using our notation above, we are interested in $P(\bar{X} \leq 81.3)$ and $P(\bar{Y} \leq 81.3)$. By the central limit theorem, we know that,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{200}\right) = N\left(80, \frac{144}{200}\right)$$
$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{200}\right) = N\left(120, \frac{121}{200}\right)$$

Using the `pnorm()` function to calculate the probability that each average is smaller than 81.3, we have,

```r
# DBP Average: P(\bar{X} <= 81.3)
pnorm(81.3, 80, sqrt(144/200))
```

```
[1] 0.9372468
```

```r
# SBP Average: P(\bar{Y} <= 81.3)
pnorm(81.3, 120, sqrt(121/200))
```

```
[1] 0
```

Again, $P(\bar{Y} \leq 81.3)$ is not actually zero, but very close to zero.

Problem 5

Suppose that IQs in a particular population are normally distributed with a mean of 110 and a standard deviation of 10.

a. What's the probability that a randomly selected person from this population has an IQ between 95 and 115?

Let X_i denote the IQ of person i randomly sampled from this population. Then,

$$P(95 \leq X_i \leq 115) = P(X_i \leq 115) - P(X_i \leq 95)$$

Using the `pnorm()` function in `R`, we get that this probability is equal to,

```r
pnorm(115, 110, 10) - pnorm(95, 110, 10)
```

```
[1] 0.6246553
```

b. What's the 65^{th} percentile from this distribution?

```r
qnorm(0.65, 110, 10)
```

```
[1] 113.8532
```

c. Suppose that 5 people are sampled from this distribution. What's the probability 4 (80%) or more have IQs above 130?

Let Z_i be an indicator variable denoting whether individual i has an IQ above 130. That is,

$$Z_i = \begin{cases} 1 & \text{if } X_i > 130 \\ 0 & \text{if } X_i \leq 130 \end{cases}$$

Then, we are interested in $P\left(\sum_{i=1}^{5} Z_i \geq 4\right)$. Since the individuals are independently selected, we know that $\sum_{i=1}^{5} Z_i \sim Binomial(5, p)$ where $p = P(X_i > 130)$. Using the `pnorm()` and `pbinom()` functions in `R`,

```r
# Calculate P(X_i > 130)
p <- pnorm(130, 110, 10, lower.tail=FALSE)

# Calculate P(sum Z_i >= 4) = P(sum Z_i > 3)
pbinom(3, 5, p, lower.tail=FALSE)
```

```
[1] 1.315009e-06
```

d. Suppose that 500 people are sampled from this distribution. What's the probability 400 (80%) or more have IQs above 130?

Using the same approach as above in part (c), we are interested in $P\left(\sum_{i=1}^{500} Z_i \geq 400\right)$ where $\sum_{i=1}^{500} Z_i \sim Binomial(500, p)$ and $p = P(X_i > 130)$.

```
# Calculate P(sum Z_i >= 400) = P(sum Z_i > 399)
pbinom(399,500,p,lower.tail=FALSE)
```

```
[1] 0
```

e. Consider the average of 100 people drawn from this distribution. What's the probability that this mean is larger than 112.5?

By the CLT, we know that $\bar{X} \sim N\left(110, \frac{10^2}{100}\right)$. Thus,

```
# Calculate P(\bar{X} > 112.5)
pnorm(112.5, 110, sqrt(10^2/100), lower.tail = FALSE)
```

```
[1] 0.006209665
```

Problem 6

Suppose that 400 observations are drawn at random from a distribution with mean 0 and standard deviation 40.

a. What's the approximate probability of getting a sample mean larger than 3.5?

From the CLT, $\bar{X} \sim N\left(\mu = 0, \sigma = \frac{40}{\sqrt{400}}\right)$.

```
pnorm(3.5, 0, 40/sqrt(400), lower.tail = FALSE)
```

```
[1] 0.04005916
```

b. Was normality of the underlying distribution required for this calculation?

No – the CLT states that for large samples, the sample mean is normally distributed, regardless of the underlying population distribution.

Problem 7

Recall that R's function `runif()` generates (by default) random uniform variables that have means 1/2 and variance 1/12.

a. Sample 1,000 observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

```
set.seed(11111)
x <- runif(1000)
mean(x)
```

```
[1] 0.5024772
```

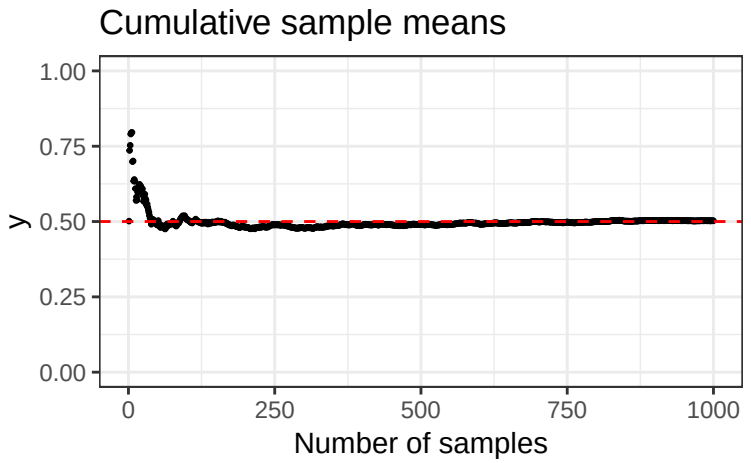
```
var(x)
```

```
[1] 0.08450761
```

The sample mean and variance should estimate the population mean and variance, respectively, since the sample mean and variance are consistent estimators for the population mean and variance.

b. Retain the same 1,000 observations from part a. Plot the sequential sample means by observation number. Hint. If x is a vector containing the simulated uniforms, then the code `y <- cumsum(x) / (1 : length(x))` will create a vector of the sequential sample means. Explain the resulting plot.

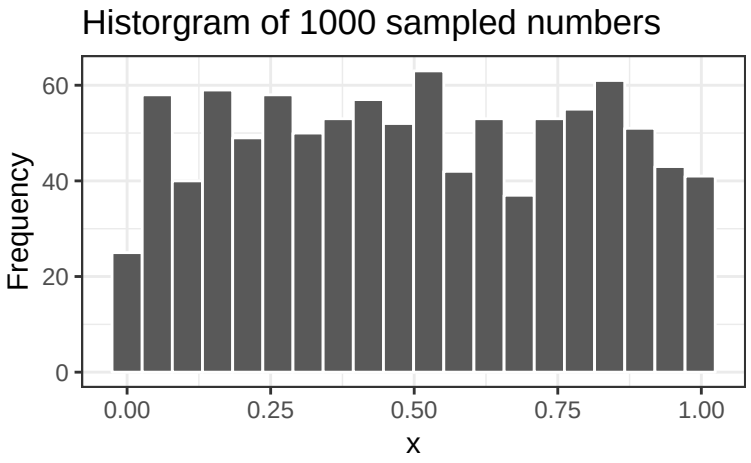
```
y <- cumsum(x)/(1:length(x))

ggplot(data = data.frame(n.samples = 1:1000, y = y)) +
  geom_point(aes(n.samples, y), size = 0.5) +
  geom_hline(yintercept = 0.5, color = 'red', linetype = 'dashed') +
  xlim(0,1000) + ylim(0, 1) +
  labs(title = "Cumulative sample means",
       x = "Number of samples", "y = Sample mean") + theme_bw()
```



The plot shows that when the sample size gets larger, the sample mean becomes closer to the population mean, i.e. the sample mean converges to 0.5. This is an illustration of the Law of Large Numbers.

c. Plot a histogram of the 1,000 numbers. Does it look like a uniform density?

```
ggplot(data.frame(sample = 1:1000, x = x)) +
  geom_histogram(aes(x), bins = 20, color = 'white') +
  labs(title = "Historgram of 1000 sampled numbers",
       x = "x", y = "Frequency") + theme_bw()
```



Yes, the density looks fairly uniform.

d. Now sample 1,000 *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these 1,000 numbers be equal to and why? Hint. The command

```
x <- matrix(runif(1000 * 100), nrow = 1000)
```

creates a matrix of size 1,000×100 filled with random uniforms. The command `y <- apply(x,1,mean)` takes the sample mean of each row.

```
# Sample 100 uniform r.v. 1000 times each
x <- matrix(runif(1000*100), nrow=1000)
# Derive the sample mean for each of the 100 uniform r.v.
y <- apply(x, 1, mean)
# Calculate mean and variance of sample means
mean(y)
```

```
[1] 0.4989176
```

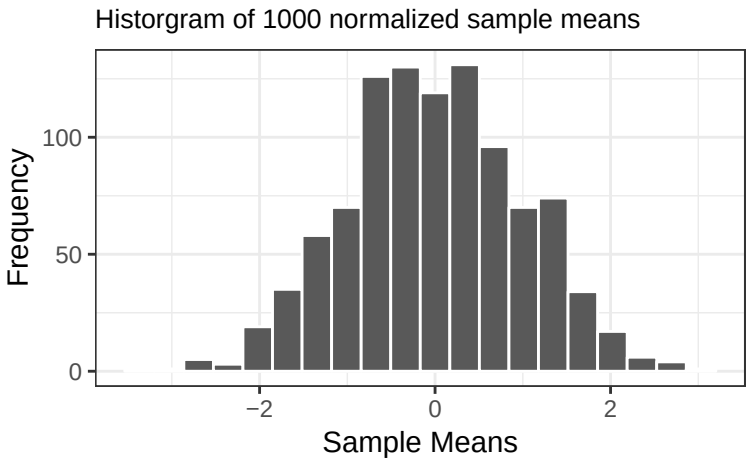
```
var(y)
```

```
[1] 0.0008390581
```

Let μ_U, σ_U^2 denote the mean and variance of the uniform distribution. The central limit theorem says that the sample mean of 100 samples from the uniform follows a normal distribution with mean and variance $\mu_U, \sigma_U^2/100$, respectively. Since the sample mean and variance are consistent estimators for the population mean and variance, we have that the sample mean of y and the sample variance y should estimate the population mean and variance of y, μ_U and $\sigma_U^2/100$, respectively. These numbers indeed match.

e. Plot a histogram of the 1,000 sample means appropriately normalized. What does it look like and why?

```
y.norm <- (y - mean(y))/(sd(y)) # Normalize sample means
ggplot(data.frame(y.norm)) +
  geom_histogram(aes(y.norm), bins = 20, color = 'white') +
  labs(title = "Historgram of 1000 normalized sample means",
       x = "Sample Means", y = "Frequency") + theme_bw()  +
  theme(plot.title = element_text(size = 10))
```



This graph looks normal because according to the central limit theorem, the distribution of sample means is approximately normal.

f. Now sample 1,000 *sample variances* from this distribution, each comprised of 100 observations. Take the average of these 1,000 variances. What property does this illustrate and why?

```
# Derive the sample variance for each of the 100 uniform r.v samples
z <- apply(x, 1, var)
```

```r
# Calculate the mean of the sample variances
mean(z)
```

```
[1] 0.08352524
```

The mean of the sample variances should converge to the true population variance. This illustrates the property of consistency of sample estimators.

Problem 8

Note that R's function `rexp` generates random exponential variables. The exponential distribution with rate 1 (the default) has a theoretical mean of 1 and variance of 1.

a. Sample 1,000 observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

```r
set.seed(22222)
x <-rexp(1000)
mean(x)
```

```
[1] 0.9484545
```

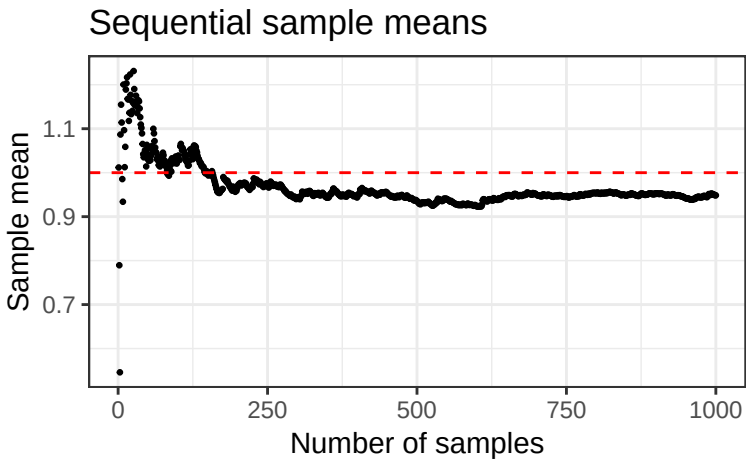
```r
var(x)
```

```
[1] 0.800305
```

By the Law of Large Numbers, sample means and sample variances taken from a population should approximate the population mean and variance when the sample size is sufficiently large.

b. Retain the same 1,000 observations from part a. Plot the sequential sample means by observation number. Explain the resulting plot.

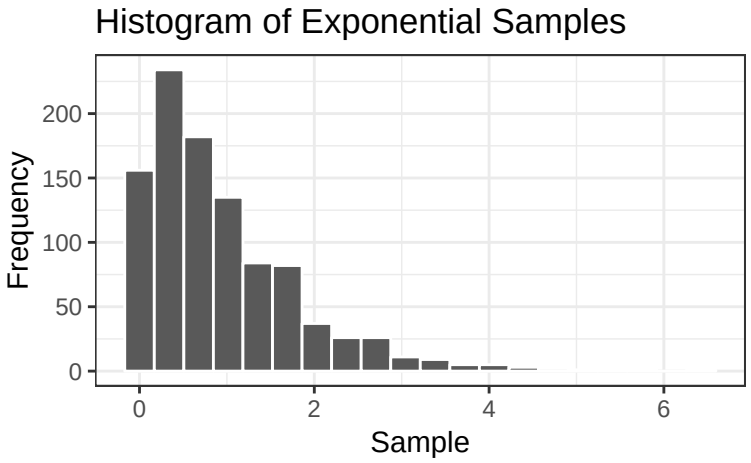
```r
# Calculate sequential averages
xbar.seq <- cumsum(x)/(1:1000)
# Plot
ggplot(data.frame(x = 1:1000, xbar.seq)) +
  geom_point(aes(x = x, y = xbar.seq), size = 0.5) +
  geom_hline(yintercept = 1, color = 'red', linetype = 'dashed')+
  labs(title = "Sequential sample means",
       x = "Number of samples", y = "Sample mean") +
  theme_bw()
```



The plot shows that when the sample size gets larger, the sample mean become closer to the population mean of 1. This is illustration of the Law of Large Numbers.

c. Plot a histogram of the 1,000 numbers. Does it look like a exponential density?

```
ggplot(data.frame(sample = x)) +
  geom_histogram(aes(x = sample), bins = 20, color = 'white') +
  labs(title = "Histogram of Exponential Samples",
       x = "Sample", y = "Frequency") +
  theme_bw()
```



Yes, the histogram does look like an exponential density.

d. Now sample 1,000 *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these 1,000 numbers be equal to and why?

```
# Create matrix of 1000 trials, each with 100 observations
x <- matrix(rexp(1000*100), nrow = 1000)
# Take mean of each trial
s.means <- apply(x, 1, mean)
# Calculate mean and variance of each sample mean
mean(s.means)
```

```
[1] 0.9971158
```

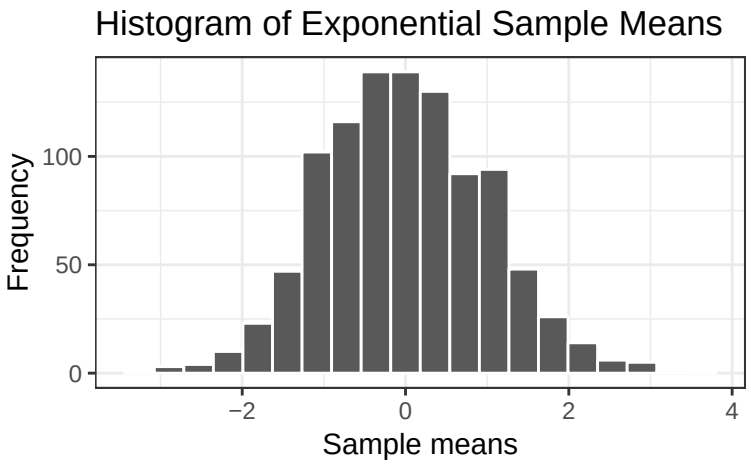
```
var(s.means)
```

```
[1] 0.009493402
```

Let μ_E, σ_E^2 denote the mean and variance of the exponential distribution. The central limit theorem says that the sample mean of 100 samples from the exponential follows a normal distribution with mean and variance $\mu_E, \sigma_E^2/100$, respectively. Since the sample mean and variance are consistent estimators for the population mean and variance, we have that the sample mean of `s.means` and the sample variance of `s.means` should estimate the population mean and variance of `s.means`, μ_E and $\sigma_E^2/100$, respectively. These numbers indeed match.

e. Plot a histogram of the 1,000 sample means appropriately normalized. What does it look like and why?

```
s.means_norm <- (s.means - mean(s.means))/sd(s.means)
ggplot(data.frame(s.means_norm)) +
  geom_histogram(aes(x = s.means_norm), bins = 20, color = 'white') +
  labs(title = "Histogram of Exponential Sample Means",
       x = "Sample means", y = "Frequency") +
  theme_bw()
```



The distribution of normalized sample means looks normal and centered at the population mean. According to the central limit theorem, the distribution of sample means should be approximately normal.

f. Now sample 1,000 *sample variances* from this distribution, each comprised of 100 observations. Take the average of these 1,000 variances. What property does this illustrate and why?

```
s.var <- apply(x, 1, var)
mean(s.var)
```

```
[1] 0.9894342
```

The mean of the sample variances should converge to the true population variance. This illustrates the property of consistency of sample estimators.

Problem 9

Consider the distribution of a fair coin flip (i.e. a random variable that takes the values 0 and 1 with probability 1/2 each.)

a. Sample 1,000 observations from this distribution. Take the sample mean and sample variance. What numbers should these estimate and why?

```
set.seed(3333)
# Sample 1000 coin tosses
coins <- rbinom(1000, 1, 0.5)
# Calculate Mean and Variance
mean(coins)
```

```
[1] 0.487
```

```
var(coins)
```

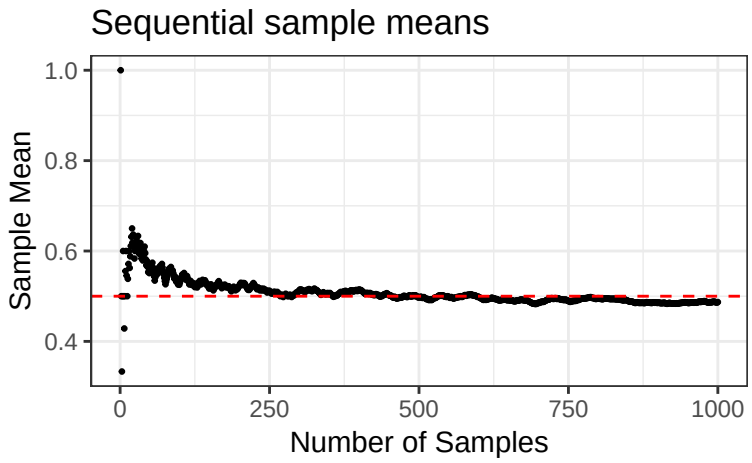
```
[1] 0.2500811
```

The sample mean and variance should approximate the population mean and variance of the bernoulli distribution of \$p = 0.5 \$ and $p(1-p) = 0.5 * 0.5 = 0.25$, respectively.

b. Retain the same 1,000 observations from part a. Plot the sequential sample means by observation number. Explain the resulting plot.

```
coins.seq <- cumsum(coins)/(1:1000)

ggplot(data.frame(coins.seq)) +
```

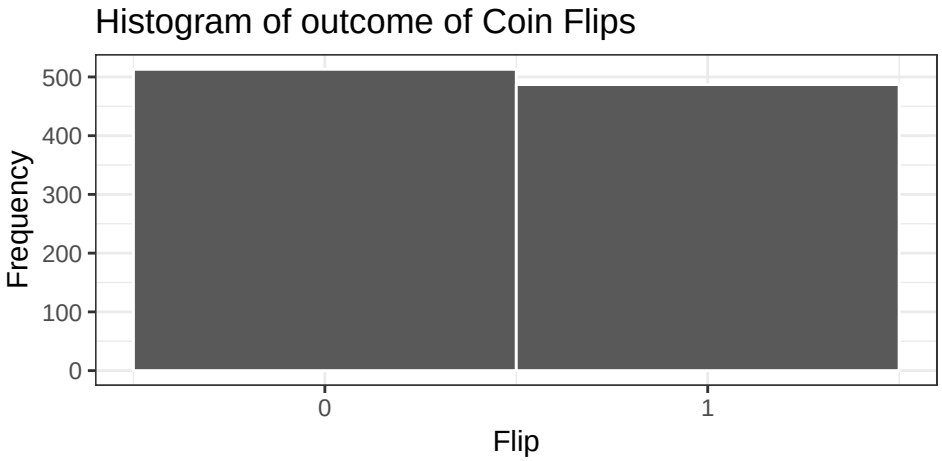
```
  geom_point(aes(x = 1:1000, y = coins.seq), size = 0.5) +
  geom_hline(yintercept = 0.5, color = 'red', linetype = 'dashed')+
  labs(title = "Sequential sample means",
       x = "Number of Samples", y = "Sample Mean") +
  theme_bw()
```



We see that as the number of samples increases, the sample mean converges to the true population mean of 0.5.

Plot a histogram of the 1,000 numbers. Does it look like it places equal probability on 0 and 1?

```
ggplot(data.frame(coins), aes(x = coins)) +
  geom_histogram(binwidth = 1, color = 'white') +
  labs(title = "Histogram of outcome of Coin Flips",
       x = "Flip", y = "Frequency") +
  scale_x_continuous(breaks = c(0,1)) +
  theme_bw()
```



Yes, it looks like equal weight is placed on both 0 and 1.

d. Now sample 1,000 *sample means* from this distribution, each comprised of 100 observations. What numbers should the average and variance of these 1,000 numbers be equal to and why?

```
# Create 1000 trials each with 100 coin flips
coins.x <- matrix(rbinom(100*1000, 1, 0.5), nrow = 1000)
# Calculate sample mean for each trial
```

```
coins.smean <- apply(coins.x, 1, mean)
# Calculate mean and variance of the sample mean
mean(coins.smean)
```

```
[1] 0.50162
```

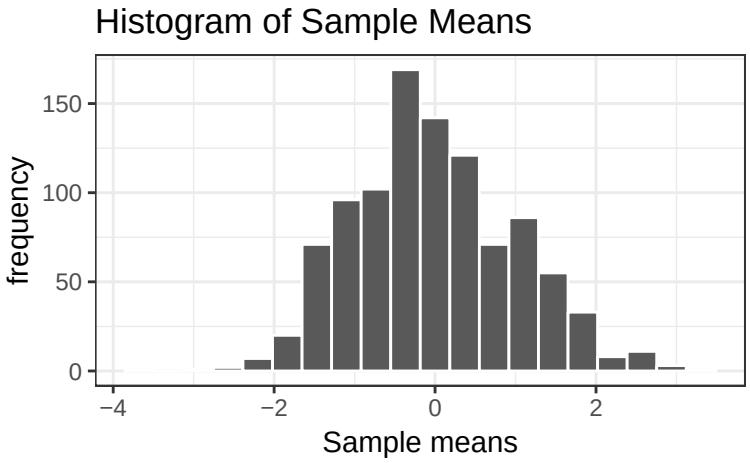
```
var(coins.smean)
```

```
[1] 0.002661637
```

Let $\mu_B = 0.5, \sigma_B^2 = 0.5 * (1 - 0.5) = 0.25$ denote the mean and the variance of the Bernoulli distribution. By the CLT, the sample mean converges to a normal distribution with mean μ_B and variance $\sigma_B^2/100$. Since the sample mean and variance approximate the population mean and variance, we expect the sample means to be approximately equal to μ_B and variance $\sigma_B^2/100$.

e. Plot a histogram of the 1,000 sample means appropriately normalized. What does it look like and why?

```
coins.norm <- (coins.smean - mean(coins.smean))/sd(coins.smean)
ggplot(data.frame(coins.norm), aes(x = coins.norm)) +
  geom_histogram(bins = 20, color = 'white') +
  labs(title = "Histogram of Sample Means",
       x = "Sample means", y = "frequency") + theme_bw()
```



By the CLT, we expect the histogram of sample means to look approximately normal.

f. Now sample 1,000 *sample variances* from this distribution, each comprised of 100 observations. Take the average of these 1,000 variances. What property does this illustrate and why?

```
coins.svar <- apply(coins.x, 1, var)
mean(coins.svar)
```

```
[1] 0.2498368
```

The expected value of the sample variance is the true population variance, i.e. the sample variance is a consistent estimator of the true population variance.

Problem 10

Consider a density for the proportion of a person's body that is covered in freckles, X, given by $f(x) = cx$ for $0 \leq x \leq 1$ and some constant c.

a. What value of c makes this function a valid density?

To be a valid density, the $\int_0^1 f(x)dx = 1$.

$$\int_0^1 f(x)dx = \int_0^1 cxdx = \frac{c}{2}x^2\Big|_0^1 = \frac{c}{2}$$

Setting the interal equal to 1 and solving for c, we get $c = 2$.

b. What is the mean and variance of this density?

$$E[X] = \int_0^1 x \cdot 2xdx = \int_0^1 2x^2dx = \frac{2}{3}x^3\Big|_0^1 = \frac{2}{3}$$

$$E[X^2] = \int_0^1 x^2 \cdot 2xdx = \int_0^1 2x^3dx = \frac{1}{2}x^4\Big|_0^1 = \frac{1}{2}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

Thus, the mean and variance of this density are $\frac{2}{3}$ and $\frac{1}{18}$, respectively.

c. You simulated 100,000 sample means, each comprised of 100 draws from this density. You then took the variance of those 100,000 numbers. Approximately what number did you obtain? (Explain.)

By the CLT, we know that the sample means, each comprised of 100 draws from the described density, are normally distributed with mean $\frac{2}{3}$ and variance $\frac{1}{18*100}$. Taking the variance of the 100,000 sample means should give us an approximation of the true variance of the sample means, $\frac{1}{18*100} = \frac{1}{1800}$.

Problem 11

Suppose that DBPs drawn from a certain population are normally distributed with a mean of 90 mmHg and standard deviation of 5 mmHg. Suppose that 1,000 people are drawn from this population.

a. If you had to guess the number of people people in having DBPs less than 80 mmHg what would you guess?

Let X denote the DBP of an individual. The probability of one person having DBP < 80 is given by $P(X < 80)$. Using the `pnorm()` function, we find that this probability is,

```
pnorm(80, 90, 5)
```

```
[1] 0.02275013
```

Thus, assuming the individuals are independently sampled, we would expect out of 1,000 individuals, 22 to 23 individuals with DBP less than 80mmHg.

b. Consider the setting for the previous problem. You draw 25 people from this population. What's the probability that the sample average is larger than 92 mmHg?

By the CLT, the sample average for 25 people follows a normal distribution with mean 90 and variance $\frac{5^2}{25} = 1$. Thus, the probability that the sample average is larger than 92 mmH is given by,

```
pnorm(92, 90, 1, lower.tail=FALSE)
```

```
[1] 0.02275013
```

c. You select 5 people from this population. What's the probability that 4 or more of them have a DBP larger than 100 mmHg?

Let $X_1, X_2, \cdots, X_5$ denote the DBP of the five individuals sampled from the population. Additionally, let

$$Z_i = \begin{cases} 1 & \text{if } X_i > 100 \\ 0 & \text{if } X_i \leq 100 \end{cases}$$

That is, Z_i is an indicator for whether individual has DBP larger than 100 mmHg. Note that $P(Z_i = 1) = P(X_i > 100)$. We are interested in the probability that 4 or more of the 5 individuals have DBP larger than 100 mmHg, or equivalently, $P\left(\sum_{i=1}^{5} Z_i \geq 4\right)$. Since the individuals are independent, we know that

$$\sum_{i=1}^{5} Z_i \sim Binomial(5, p) \qquad \text{where} p = P(Z_i = 1) = P(X_i > 100)$$

Using `R` to compute the probability, we get

```
# Calculate P(Z_i = 1) = P(X_i > 100)
p <- pnorm(100, 90, 5, lower.tail = FALSE)

# Calculate P(Sum Z_i >= 4)
pbinom(3, 5, p, lower.tail = FALSE)
```

```
[1] 1.315009e-06
```

Problem 12

You need to calculate the probability that a *standard normal* is larger than 2.20, but have nothing available other than a regular coin. Describe how you could estimate this probability using only your coin. (Do not actually carry out the experiment, just describe how you would do it.)

The CLT states that for $X_1, X_2, \cdots, X_n$, $\bar{X} \sim N(E[X], Var(X)/n)$. Thus, to estimate the probability under a standard normal, we can generate a sequence of sample means, standardize the sample means, then look at the proportion of observations above 2.2. More precisely, I would estimate the probability as follows:

1. Flip the fair coin ten times, yielding $X_1, X_2, \cdot X_{10}$. Record the sample mean $\bar{X}$.
2. Repeat *Step 1* 1000 times such that we have 1000 observations of the sample mean, which we will denote as $\bar{X}_1, \bar{X}_2, \cdots, X_{1000}$.
3. Given that X denotes the outcome of a fair coin flip, $E[X] = p = 0.5$ and $Var(X) = p(1-p) = 0.5(1-0.5) = 0.25$. By the CLT, we know that $\bar{X} \sim N(E[X], Var(X)/10) = N(0.5, 0.025)$. Thus, to get a standard normal distribution, we can normalize the sample mean by subtracting the mean and dividing by the standard deviation. That is, for each sample mean, we can derive

$$Z_i = \frac{\bar{X}_i - E[X]}{\sqrt{Var(X)/n}} = \frac{\bar{X}_i - 0.5}{\sqrt{0.025}}$$

   where $Z \sim N(0, 1)$.
4. Out of the 1000 Z_i, we can estimate the probability that a standard normal is greater than 2.2 by identifying the proportion of Z_i's that are greater than 2.2.

Problem 13

Let X_1, X_2 be independent, identically distributed coin flips (taking values 0 = failure or 1 = success) having success probability π. Give and interpret the likelihood ratio comparing the hypothesis that $\pi = .5$ (the coin is fair) versus $\pi = 1$ (the coin always gives successes) when both coin flips result in successes.

Let x be the total number of heads flipped out of 2 coins. Then the likelihood of 2 successes, each with probability π, is given by,

$$f(x,\pi) = \prod_{i=1}^{2} P(X_i = x_i) = \pi^2(1-\pi)^{2-x}$$

Assuming that both coin flips resulted in success, the likelihood ratio is given by,

$$\frac{f(x=2,\pi=0.5)}{f(x=2,\pi=1)} = \frac{(0.5)^2(1-0.5)^{2-2}}{1^2(1-1)^{2-2}} = \frac{0.5^2}{1^2} = 0.25$$

Thus, the likelihood that the coin is not fair is 4 times more supported by the data than the likelihood that the coin is fair.

Problem 14

The density for the population of increases in wages for assistant professors being promoted to associates (1 = no increase, 2 = salary has doubled) is uniform on the range from 1 to 2.

a. What's the mean and variance of this density?

The uniform density is given by,

$$f(X) = 1 \qquad \text{for } 1 \leq x \leq 2$$

Thus, the mean and variance are given by,

$$E[X] = \int_1^2 x * 1dx = \frac{1}{2}x^2\Big|_1^2 = 2 - \frac{1}{2} = \frac{3}{2}$$

$$E[X]^2 = \int_1^2 x^2 * 1dx = \frac{1}{3}x^3\Big|_1^2 = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{7}{3} - \left(\frac{3}{2}\right)^2 = \frac{1}{12}$$

b. Suppose that the sample variance of 10 observations from this density was sampled say 10,000 times. What number would we expect the average value from these 10,000 variances to be near? (Explain your answer briefly.)

By the law of large numbers, the average values should be around 1/12 (or .08) as the sample mean of the variances should converge onto the population variance.

Problem 15

Suppose that the US intelligence quotients (IQs) are normally distributed with mean 100 and standard deviation 16.

a. What IQ score represents the 5^{th} percentile? (Explain your calculation.)

Let X be a random variable for the IQ score. We are interested in the value x such that $P(X \leq x) = 0.05$, or equivalently the value such that 5% of the population scores at or below. Using the `qnorm()` function, we find, $x = 73.68$.

```
qnorm(.05,100,16)
```

```
[1] 73.68234
```

b. Consider the previous question. Note that 116 is the 84^{th} percentile from this distribution. Suppose now that 1,000 subjects are drawn at random from this population. Use the central limit theorem to write the probability that less than 82% of the sample has an IQ below 116 as a standard normal probability. Note, you do not need to solve for the final number. (Show your work.)

Again, let X_i denote the IQ score of individual i and Y_i be an indicator value for whether score $X_i < 116$. That is,

$$Y_i = \begin{cases} 1 & \text{if } X_i < 116 \\ 0 & \text{if } X_i \geq 116 \end{cases}$$

Then, $Y_i \sim Bernoulli(p)$ where $p = P(X_i < 116) = 0.84$ since 116 denotes the 84th percentile. From previous exercises, we know that $E[Y] = p$ and $var(Y) = p(1-p)$. We are interested in,

$$P\left(\sum_{i=1}^{1000} Y_i < 820\right) = P\left(\bar{Y} < 0.82\right)$$

From the CLT, we know that $\bar{Y} \sim N(E[Y], Var(Y)/1000)$. Expressing this as a standard normal, we get

$$P(\bar{Y} < 0.82) = P\left(\frac{\bar{Y} - E[Y]}{\sqrt{Var(Y)/1000}} < \frac{0.82 - E[Y]}{\sqrt{Var(Y)/1000}}\right) = P\left(\frac{\bar{Y} - 0.84}{\sqrt{0.13/1000}} < \frac{0.82 - 0.84}{\sqrt{0.13/1000}}\right) = \Phi\left(-1.725\right)$$

```
pnorm(-1.725)
```

```
[1] 0.04226374
```

c. Consider the previous two questions. Suppose now that a sample of 100 subjects are drawn from a *new* population and that 60 of the sampled subjects had an IQs below 116. Give a 95% confidence interval estimate of the true probability of drawing a subject from this population with an IQ below 116. Does this proportion appear to be different than the 84% for the population from questions 1 and 2?

Recall that the 95% confidence interval is given by,

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Given that 60 of the sampled subjects had IQs below 116, we estimate p to be $\hat{p} = \frac{60}{100}$.

```
0.6 - qnorm(1-0.05/2)*sqrt(0.6*(1-0.6)/100)    # Lowerbound
```

```
[1] 0.5039818
```

```
0.6 + qnorm(1-0.05/2)*sqrt(0.6*(1-0.6)/100)    # Upperbound
```

```
[1] 0.6960182
```

Thus, the confidence interval interval is,

$$P\left(0.6 - z_{1-\alpha/2}\sqrt{\frac{0.6*(1-0.6)}{100}}, 0.6 - z_{1-\alpha/2}\sqrt{\frac{0.6*(1-0.6)}{100}}\right) = (0.504, 0.696)$$

This proportion does appear to be different than the 84% for the population, as 0.84 is not contained within the confidence interval.

Problem 16

Let X be binomial with success probability p_1 and n_1 trials and Y be an independent binomial with success probability p_2 and n_2 trials. Let $\hat{p}_1 = X/n_1$ and $\hat{p}_2 = Y/n_2$ be the associated sample proportions. What would be an estimate for the standard error for $\hat{p}_1 - \hat{p}_2$? To have consistent notation with the next problem, label this value $\hat{SE}_{\hat{p}_1 - \hat{p}_2}$.

$$\begin{aligned}
Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(\hat{p}_2) && \text{by independence of } X \text{ and } Y \\
&= Var\left(\frac{X}{n_1}\right) + Var\left(\frac{Y}{n_2}\right) && \text{substituting for } \hat{p}_1 \text{ and } \hat{p}_2 \\
&= \frac{Var(X)}{n_1^2} + \frac{Var(Y)}{n_2^2} && \text{by properties of variance} \\
&= \frac{n_1 p_1(1-p_1)}{n_1^2} + \frac{n_2 p_2(1-p_2)}{n_2^2} && \text{since } Var(X) = p_1(1-p_1) \\
&= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}
\end{aligned}$$

To get the estimated standard error, we can substitute our estimated values $\hat{p_1}$ and $\hat{p_2}$ for p_1 and p_2, yielding

$$\hat{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Problem 17

You are in desperate need to simulate standard normal random variables yet do not have a computer available. You do, however, have ten standard six sided dice. Knowing that the mean of a single die roll is 3.5 and the standard deviation is 1.71, describe how you could use the dice to approximately simulate standard normal random variables. (Be precise.)

The CLT states that for $X_1, X_2, \cdots, X_n$, $\bar{X} \sim N(E[X], Var(X)/n)$. Thus, to simulate standard normal variables, we can generate a sequence of sample means then standardize the sample means. More precisely, we could generate standard normal random variables as follows:

1. Roll the 10 dice, yielding $X_1, X_2, \cdot X_{10}$. Record the sample mean $\bar{X}$.
2. Repeat *Step 1* 1000 times such that we have 1000 observations of the sample mean, which we will denote as $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{1000}$.
3. Given that X denotes the outcome of a die roll and we are given, $E[X] = 3.5$ and $Var(X) = 1.71^2$. By the CLT, we know that $\bar{X} \sim N(E[X], Var(X)/10) = N(3.5, 1.71^2/10)$. Thus, to get a standard normal distribution, we can normalize the sample mean by subtracting the mean and dividing by the variance. That is, for each sample mean, we can derive

$$Z_i = \frac{\bar{X}_i - E[X]}{\sqrt{Var(X)/n}} = \frac{\bar{X}_i - 3.5}{1.71/\sqrt{10}}$$

where $Z \sim N(0, 1)$. Thus, we have generated 1000 draws from a standard normal distribution.

Problem 18

In a sample of 40 United States men contained 25% smokers. Let p be the true prevalence of smoking amongst males in the United States. Write out and draw and interpret the likelihood for p. Is $p = .35$ or $p = .15$

better supported given the data (why, and by how much)? What value of p is best supported (just give the number, do not derive)?

Let x denote the number of smokers observed and assumpe that p is the true prevalence of smoking amonst males in the US. Then, we have that $X \sim Binom(40, p)$, and the likelihood is given by,

$$L(p|x) = \binom{40}{x} p^x (1-p)^{40-x}$$

the ratio of the likelihoods for $p = 0.35$ and $p = 0.15$ assuming that we observed 25% smokers in a population of 40,

$$\frac{L(p = 0.35|x = 10)}{L(p = 0.15|x = 10)} = \frac{\binom{40}{10} 0.35^{10}(1-0.35)^{40-10}}{\binom{40}{10} 0.15^{10}(1-0.15)^{40-10}} = 1.53$$

From the likelihood ratio, we see that $p = 0.35$ is better supported by the data because it has the larger likelihood when compared with $p = 0.15$. In fact, $p = 0.35$ is 1.53 times more supported by the data than $p = 0.15$.

The best supported value of p is given by 0.25 as the MLE for p in a Binomial distribution is the proportion of individuals who smoke.

Problem 19

Consider three sample variances, S_1^2, S_2^2 and S_3^2. Suppose that the sample variances are comprised of n_1, n_2 and n_3 iid draws from normal populations $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ and $N(\mu_3, \sigma^2)$, respectively. Argue that

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-3}$$

is an unbiased estimate of σ^2.

Let

$$\hat{\sigma}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-3}$$

To show that $\hat{\sigma}^2$ is an unbiased estimate of σ^2, we want to show that $E[\hat{\sigma^2}] = \sigma^2$.

$$\begin{aligned}
E\left[\hat{\sigma}^2\right] &= E\left[\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-3}\right] \\
&= \frac{1}{n_1+n_2+n_3-3} E\left[(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2\right] \\
&= \frac{1}{n_1+n_2+n_3-3} \left((n_1-1)E[S_1^2] + (n_2-1)E[S_2^2] + (n_3-1)E[S_3^2]\right) \\
&= \frac{1}{n_1+n_2+n_3-3} \left((n_1-1)\sigma^2 + (n_2-1)\sigma^2 + (n_3-1)\sigma^2\right) \\
&= \frac{1}{n_1+n_2+n_3-3} (n_1+n_2+n_3-3)\sigma^2 = \sigma^2
\end{aligned}$$

Thus, $\hat{\sigma}^2$ is an unbiased estimate of σ^2.

Problem 20

You need to calculate the probability that a normally distributed random variable is less than 1.25 standard deviations below the mean. However, you only have an oddly shaped coin with a known probability of heads

of 0.6. Describe how you could estimate this probability using this coin. (Do not actually carry out the experiment, just describe how you would do it.)

The CLT states that for $X_1, X_2, \cdots, X_n$, $\bar{X} \sim N(E[X], Var(X)/n)$. Thus, to estimate the probability that a normal variable is less than 1.25 standard deviations below the mean, we can generate a sequence of sample means, standardize the sample means, then look at the proportion of observations below -1.25. More precisely, I would estimate the probability as follows:

1. Flip the coin ten times, yielding $X_1, X_2, \cdot X_{10}$. Record the sample mean $\bar{X}$.
2. Repeat *Step 1* 1000 times such that we have 1000 observations of the sample mean, which we will denote as $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{1000}$.
3. Given that X denotes the outcome of a coin flip, $E[X] = p = 0.6$ and $Var(X) = p(1-p) = 0.6(1-0.6) = 0.24$. By the CLT, we know that $\bar{X} \sim N(E[X], Var(X)/10) = N(0.6, 0.024)$. Thus, to get a standard normal distribution, we can normalize the sample mean by subtracting the mean and dividing by the standard deviation. That is, for each sample mean, we can derive
$$Z_i = \frac{\bar{X}_i - E[X]}{\sqrt{Var(X)/n}} = \frac{\bar{X}_i - 0.6}{\sqrt{0.024}}$$
where $Z \sim N(0, 1)$.
4. Out of the 1000 Z_i, we can estimate the probability that a standard normal is less than 1.25 standard deviations below the mean by identifying the proportion of Z_i's that are less than -1.25.

There are some limitations with this approach. The accuracy is limited when the number of coin flips in step 1 is small and when there are few sample means. To calculate the most accurate answer, choose a large sample size, and generate many sample means.

Problem 21

The next three questions (A., B., C.) deal with the following setting. Forced expiratory volume, FEV_1, is a measure of lung function that is often expressed as a proportion of lung capacity called forced vital capacity, FVC. Suppose that the population distribution of FEV_1/FVC of asthmatics adults in the US has mean of .55 and standard deviation of .10.

a. Suppose a random sample of 100 people are drawn from this population. What is the probability that their average FEV_1/FVC is larger than .565?

Let $X_1, X_2, \cdots, X_{100}$ denote the FEV_1/FVC of 100 individuals sampled from the population of asthmatic adults. By the CLT, we know that $\bar{X} \sim N(E[X], Var(X)/100) = N(0.55, 0.1^2/100)$. Thus, the probability that the average is larger than 0.565 is given by $P(\bar{X} > 0.565)$. Using the `pnorm()` function, we get

```
# P(x.bar > 0.565)
pnorm(0.565, 0.55, 0.1/sqrt(100), lower.tail = FALSE)
```

```
[1] 0.0668072
```

b. Suppose the population of non-asthmatics adults in the US have a mean FEV_1/FVC of 0.8 and a standard deviation of 0.05. You sample 100 people from the asthmatic population and 100 people from the non-asthmatic population and take the difference in sample means. You repeat this process 10,000 times to obtain 10,000 differences in sample means. What would you guess the mean and standard deviation of these 10,000 numbers would be?

Let $\bar{X}_{na}$ and $\bar{X}_a$ denote the mean of the 10,000 sample means for non-asthmatic, and asthmatic adults, respectively. The expected value of the sample mean is equal to the population mean when there are many sample means. Thus, my best guess for the difference in the FEV_1/FVC of non-asthmatic adults and asthmatic adults would be,

$$E[\bar{X}_{na} - \bar{X}] = E[\bar{X}_{na}] - E[\bar{X}_a] = \mu_{na} - \mu_a = 0.80 - 0.55 = 0.25$$

meaning the sample mean of the non asthmatics is on average 0.25 higher than the sample mean of asthmatics.

The variance of the difference in FEV_1/FVC between non-asthmatic and asthmatic adults assuming independence between the two populations is,

$$\begin{aligned} Var(\bar{X}_{na} - \bar{X}_a) &= Var(\bar{X}_{na}) + Var(\bar{X}_a) && \text{by independence} \\ &= \frac{\sigma^2_{na}}{n_{na}} + \frac{\sigma^2_a}{n_a} && \text{by CLT} \\ &= \frac{0.05^2}{100} + \frac{0.1^2}{100} = 0.000125 \end{aligned}$$

Thus, the standard deviation of the difference is given by $\sqrt{Var(\bar{X}_{na} - \bar{X}_a)} = \sqrt{0.000125} = 0.0112$.

c. Moderate or severe lung dysfunction is defined as $FEV_1/FVC \leq .40$. A colleague tells you that 60% of asthmatics in the US have moderate or severe lung dysfunction. To verify this, you take a random sample of 5 subjects, only one of which has moderate or severe lung dysfunction. What is the probability of obtaining only one or fewer if your friend's assertion is correct? What does your result suggest about their assertion?

Let Z_i be an indicator for whether individual i has moderate or severe lung dysfunction. That is, using our notation above,

$$Z_i = \begin{cases} 1 & \text{if } X_i \leq 0.4 \quad 0.6 \\ 0 & \text{if } X_i > 0.4 \quad 0.4 \end{cases}$$

Since we assume independence of individuals, we have that $\sum_{i=1}^5 Z_i \sim Binomial(n, p)$ where \$p is the probability of moderate or severe lung dysfunction. Thus, we are interested in,

$$P\left(\sum_{i=1}^5 Z_i \leq 1\right) = P\left(\sum_{i=1}^5 Z_i = 1\right) + P\left(\sum_{i=1}^5 Z_i = 0\right) = \binom{5}{1} p^1 (1-p)^{5-1} + \binom{5}{0} p^0 (1-p)^0$$

Assuming that my friend is correct, $p = 0.6$. Using the `pbinom()` function, we get that the probability of seeing one or fewer individuals with severe or moderate lung dysfunction is,

```
pbinom(1, 5, 0.6)
```

```
[1] 0.08704
```

Thus, obtaining the result of 1 or 0 persons with severe lung dysfunction out of 5 in a random sample is unlikely given a probability of 0.6. This suggests either that the probability of dysfunction is actually much lower or that our sample was not representative of the total population.

Problem 22

Consider a sample of n iid draws from an exponential density

$$\frac{1}{\beta} \exp(-x/\beta) \quad \text{for} \quad \beta > 0$$

a. Derive the maximum likelihood estimate for β.

Let $\mathbf{x} = x_1, x_2, \cdots, x_n$ be the iid draws from an exponential density. Deriving the MLE for β,

$$L(\beta|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\beta} \exp(-x_i/\beta) = \frac{1}{\beta^n} \exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right\}$$

$$\log L(\beta|\mathbf{x}) = -n\log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} x_i$$

$$\frac{d}{d\beta}\log L(\beta|\mathbf{x}) = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i$$

Setting the derivative equal to 0 and solving for β, we get

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

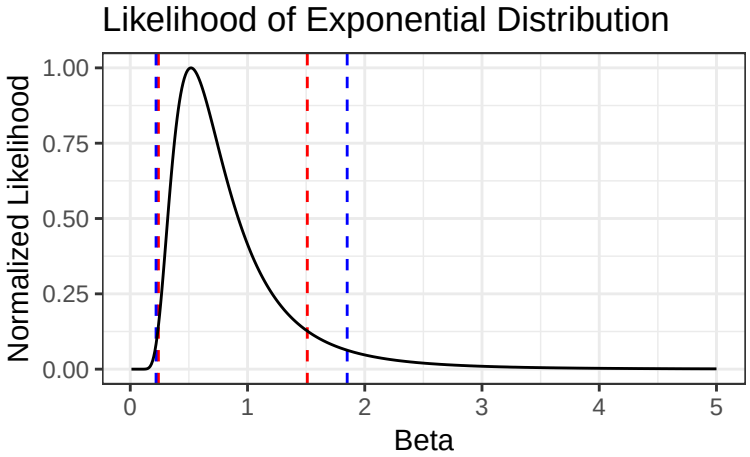
Evaluating the second derivative of $\log L(\beta|\mathbf{x})$ at $\beta = \hat{\beta}$ to verify that $\hat{\beta}$ maximizes the likelihood,

$$\frac{d^2}{d\beta^2}\log L(\beta|\mathbf{x})\bigg|_{\beta=\hat{\beta}} = \frac{n}{\beta^2} - \frac{2}{\beta^3}\sum_{i=1}^{n} x_i\bigg|_{\beta=\hat{\beta}} = \frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} - \frac{2n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} = -\frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} < 0$$

Thus $\hat{\beta}$ is indeed an MLE for β.

b. Suppose that in your experiment, you obtained five observations, `1.590 0.109 0.155 0.281 0.453`, plot the likelihood for β. Put in reference lines at 1/8 and 1/16.

```
beta <- seq(0.01, 5, by = 0.01)
data <- c(1.590, 0.109, 0.155, 0.281, 0.453)
likelihood <- 1/beta^5 * exp(-1/beta*sum(data))
likelihood.norm <- likelihood/max(likelihood)
# Identify reference lines
r8 <- beta[range(which(likelihood.norm > 1/8))]     # For 1/8
r16 <- beta[range(which(likelihood.norm > 1/16))]   # For 1/16

ggplot(data.frame(beta, likelihood.norm), aes(x = beta, y = likelihood.norm)) +
  geom_vline(xintercept = r8, color = 'red', linetype = 'dashed')+
  geom_vline(xintercept = r16, color = 'blue', linetype = 'dashed') +
  geom_line() + labs(title = "Likelihood of Exponential Distribution",
       x = "Beta", y = "Normalized Likelihood") + theme_bw()
```



Problem 23

Often infection rates per time at risk are modelled as Poisson random variables. Let X be the number of infections and let t be the person days at risk. Consider the Poisson mass function $(t\lambda)^x \exp(-t\lambda)/x!$. The parameter λ is called the population incident rate.

a. Derive the ML estimate for λ.

Deriving the MLE for λ,

$$\begin{aligned} L(\lambda|X,t) &= (t\lambda)^X \exp(-t\lambda)/X! \\ \log L(\lambda|X,t) &= X\log(t\lambda) - t\lambda - \log(X!) \\ \frac{d}{d\lambda}\log L(\lambda|X,t) &= \frac{X}{\lambda} - t \end{aligned}$$

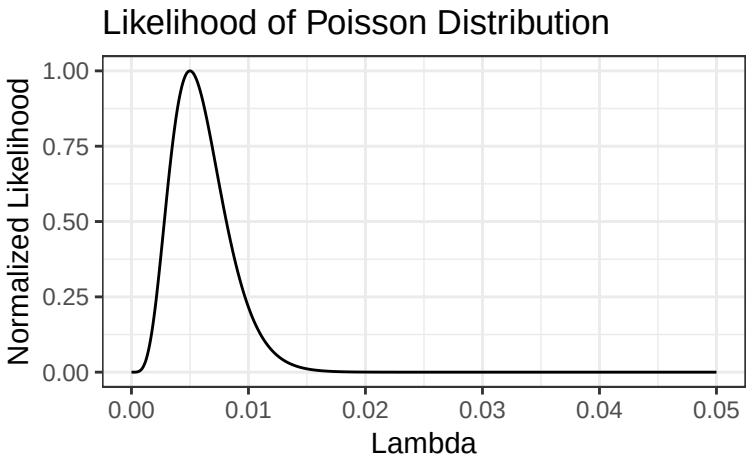
Setting the derivative equal to 0 and solving for λ, we get $\hat{\lambda} = \frac{X}{t}$. Evaluating the second derivative of $\log L(\lambda|X,t)$ at $\lambda = \hat{\lambda}$ to verify that $\hat{\lambda}$ maximizes the likelihood,

$$\frac{d^2}{d\lambda^2}\log L(\lambda|X,t) = -\frac{X}{\lambda^2} < 0$$

Thus $\hat{\lambda}$ is indeed an MLE for λ.

b. Suppose that 5 infections are recorded per 1000 person-days at risk. Plot the likelihood.

```
lambda <- seq(0, 0.05, by = 0.05/1000)
likelihood <- (1000*lambda)^5 * exp(-1000*lambda)/factorial(5)
norm.lik <- likelihood/max(likelihood)

ggplot(data.frame(lambda, norm.lik), aes(x = lambda, y = norm.lik)) +
  geom_line() + labs(title = "Likelihood of Poisson Distribution",
       x = "Lambda", y = "Normalized Likelihood") + theme_bw()
```



c. Suppose that five independent hospitals are monitored and that the infection rate (λ) is assumed to be the same at all five. Let X_i, t_i be the count of the number of infections and person days at risk for hospital i. Derive the ML estimate of λ.

Let $\mathbf{X} = (X_1, X_2, \cdots, X_5)$ and $\mathbf{t} = (t_1, t_2, \cdots, t_5)$. Then, the ML estimatie of λ is given by,

$$L(\lambda|\mathbf{t},\mathbf{X}) = \prod_{i=1}^{5} \frac{(t_i\lambda)^{X_i} e^{-t_i\lambda}}{X_i!}$$

$$\log L(\lambda|\mathbf{t},\mathbf{X}) = \sum_{i=1}^{5} (X_i \log(t_i\lambda) - t_i\lambda - \log X_i!) \propto \sum_{i=1}^{5} X_i \log(t_i\lambda) - \lambda \sum_{i=1}^{5} t_i$$

$$\frac{d}{d\lambda} \log L(\lambda|\mathbf{t},\mathbf{X}) = \sum_{i=1}^{5} \frac{X_i}{\lambda} - \sum_{i=1}^{5} t_i = 0$$

Setting the derivative equal to 0 and solving for λ, we get that the MLE for λ is,

$$\hat{\lambda} = \frac{\sum_{i=1}^{5} X_i}{\sum_{i=1}^{5} t_i}$$

To verify that $\hat{\lambda}$ is the value of λ that maximizes the likelihood, we can look at the second derivative evaluate at $\lambda = \hat{\lambda}$.

$$\frac{d^2}{d\lambda^2} \log L(\lambda|\mathbf{t},\mathbf{X}) = -\frac{\sum_{i=1}^{5} X_i}{\lambda^2}$$

Since $X_i > 0$, we have that the second derivative of the log-likelihood is indeed negative, so $\hat{\lambda}$ is the MLE.

Problem 24

Consider n iid draws from a gamma density where α is known

$$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-x/\beta) \quad \text{for} \quad \beta > 0, x > 0, \alpha > 0.$$

a. Derive the ML estimate of β.

Let $\mathbf{X} = (X_1, \cdots, X_n)$. Deriving the ML estimate of β,

$$L(\beta|\mathbf{X},\alpha) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X_i^{\alpha-1} \exp(-X_i/\beta)$$

$$\log L(\beta|\mathbf{X},\alpha) = \sum_{i=1}^{n} (-\log(\Gamma(\alpha)) - \alpha \log(\beta) - (\alpha-1)\log(X_i) - X_i/\beta)$$

$$\frac{d}{d\beta} \log L(\beta|\mathbf{X},\alpha) = \sum_{i=1}^{n} \left(-\frac{\alpha}{\beta} + \frac{X_i}{\beta^2}\right) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} X_i$$

Setting the derivative equal to zero and solving for β, we get

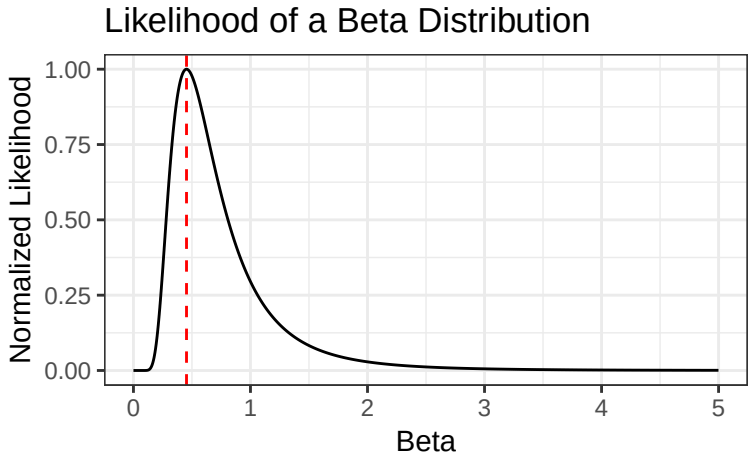
$$\hat{\beta} = \frac{1}{n\alpha} \sum_{i=1}^{n} X_i = \frac{\bar{X}}{\alpha}$$

Verifying that $\hat{\beta}$ is the MLE of β, by showing that the second derivative is negative at $\beta = \hat{\beta}$,

$$\left(\frac{d^2}{d\beta^2} \log L(\beta|\mathbf{X},\alpha)\right)\Bigg|_{\beta=\hat{\beta}} = \left(\frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} X_i\right)\Bigg|_{\beta=\hat{\beta}} = \frac{n^3\alpha^3}{\sum_{i=1}^{n} X_i} - \frac{2n^3\alpha^3}{\sum_{i=1}^{n} X_i} = -\frac{n^3\alpha^3}{\sum_{i=1}^{n} X_i} < 0$$

b. Suppose that $n = 5$ observations were obtained: 0.015, 0.962, 0.613, 0.061, 0.617. Draw a likelihood plot for β (still assume that $\alpha = 1$).

```
beta <- seq(0.0001, 5,by=0.001)
data <- c(0.015, 0.962, 0.613, 0.061, 0.617)
likelihood <- sapply(beta, function(x) prod(dgamma(data, shape = 1, scale = x)))
likelihood.norm <- likelihood/max(likelihood)

ggplot(data.frame(beta, likelihood.norm), aes(x = beta, y = likelihood.norm)) +
  geom_vline(xintercept = mean(data), color = 'red', linetype = 'dashed') +
  geom_line() + labs(title = "Likelihood of a Beta Distribution",
       x = "Beta", y = "Normalized Likelihood") + theme_bw()
```



The MLE according to part (a) should be

$$\hat{\beta} = \frac{\sum_{i=1}^{5} X_i}{n\alpha} = \frac{0.015 + 0.962 + 0.613 + 0.061 + 0.617}{5 * 1} = 0.454.$$

The plot confirms this, as the maximum of the likelihood function occurs at 0.454.

Problem 25

a. Let $Y_1, \ldots, Y_N$ be iid random variabels from a Lognormal distribution with parameters μ and σ^2. Note $Y \sim \text{Lognormal}(\mu, \sigma^2)$ if and only if $\log Y \sim N(\mu, \sigma^2)$. The log-normal density is given by

$$(2\pi\sigma^2)^{-1/2} \exp[-\{\log(y) - \mu\}^2/2\sigma^2]/y \quad \text{for} \quad y > 0$$

a. Show that the ML estimate of μ is $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \log(Y_i)$. (The mean of the log of the observations. This is called the "geometric mean".)

Let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_N)$. Deriving the MLE estimate of the mean,

$$L(\mu|\mathbf{Y}, \sigma^2) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} \exp[-\{\log(Y_i) - \mu\}^2/2\sigma^2]/Y_i$$

$$\log L(\mu|\mathbf{Y}, \sigma^2) = \sum_{i=1}^{N} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\log(Y_i) - \mu)^2}{2\sigma^2} - \log(Y_i) \right)$$

$$\frac{d}{d\mu} \log L(\mu|\mathbf{Y}, \sigma^2) = \sum_{i=1}^{N} \frac{\log(Y_i) - \mu}{\sigma^2}$$

Setting the derivative of the log-likelihood equal to zero and solving for μ,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\log(Y_i)$$

Checking that the second derivative is negative,

$$\frac{d^2}{d\mu^2}\log L(\mu|\mathbf{Y},\sigma^2) = -\frac{n}{\sigma^2} < 0$$

Thus $\hat{\mu}$ is an MLE for μ.

b. Show that the ML estimate of σ^2 is then the biased variance estimate based on the log observation

$$\frac{1}{N}\sum_{i=1}^{N}(\log(y_i) - \hat{\mu})^2$$

Using the same notation as above, and using our ML estimate for μ, we get that the ML estimate for σ^2 is,

$$\begin{aligned}
L(\sigma^2|\mathbf{Y},\hat{\mu}) &= \prod_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\exp[-\{\log(Y_i)-\hat{\mu}\}^2/2\sigma^2]/Y_i \\
\log L(\sigma^2|\mathbf{Y},\hat{\mu}) &= \sum_{i=1}^{N}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\log(Y_i)-\hat{\mu})^2}{2\sigma^2} - \log(Y_i)\right) \\
\frac{d}{d\sigma^2}\log L(\sigma^2|\mathbf{Y},\hat{\mu}) &= \sum_{i=1}^{N}\left(-\frac{1}{2\sigma^2} + \frac{(\log(Y_i)-\hat{\mu})^2}{2\sigma^4}\right) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(\log(Y_i)-\hat{\mu})^2
\end{aligned}$$

Setting the derivative of the log-likelihood equal to zero and solving for σ, we get that the ML estimate $\hat{\sigma}^2$ is,

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(\log(Y_i)-\hat{\mu})^2$$

The second derivative of the log-likelihood evaluated at $\sigma^2 = \hat{\sigma}^2$ is,

$$\begin{aligned}
\frac{d^2}{d(\sigma^2)^2}\log L(\sigma^2|\mathbf{Y},\hat{\mu})\bigg|_{\sigma^2=\hat{\sigma}^2} &= \left(\frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{N}(\log(Y_i)-\hat{\mu})^2\right)\bigg|_{\sigma^2=\hat{\sigma}^2} \\
&= \frac{N^3}{2\left(\sum_{i=1}^{N}(\log Y_i - \hat{\mu})^2\right)^2} - \frac{N^3}{\left(\sum_{i=1}^{N}(\log Y_i - \hat{\mu})^2\right)^3} \\
&= -\frac{N^3}{\left(\sum_{i=1}^{N}(\log Y_i - \hat{\mu})^2\right)^3}
\end{aligned}$$

Since the second derivative is negative at $\sigma^2 = \hat{\sigma}^2$, $\hat{\sigma}^2$ is the ML estimate of σ^2.