Homework 9. Due Friday, December 3rd. All of these questions have computational solutions. Your solution must be in a `.R` or other similar file.

1 Preamble

The first part of this homework is focused on using the Central Limit Theorem (CLT) and Slutsky's Theorem (a version of which is Theorem 8.3 in Trosset's book). There are essentially two kinds of problems in this homework. The first kind assume that the variance in the population is known. These correspond to the classical CLT. The second kind do not assume that the population variance is known. These use Slutsky's theorem to invoke the CLT using a consistent estimator of the population variance.

The second part of this homework is focused on Wald Tests. These are tests built from the CLT and use asymptotic normality under the null hypothesis to build a test statistic. There are essentially two kinds of null hypotheses (one-sided and point null hypotheses). Once a test is devised, then one can ask for the complement of the rejected null hypotheses. This set provides a route to forming an interval estimate. Because the Wald Tests are built off of the CLT, there is essentially a one-to-one connection between interval estimates from inverting the Wald test and interval estimates we would get directly from the CLT.

Load the data for the problems using the commands locally given or by loading the file `hw_9.RData` which contains the data vectors for six of the questions.

1.1 The CLT

Suppose that $X_1, \ldots, X_n, \ldots \overset{iid}{\sim} F$ where $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Let $\overline{X}_n = (X_1 + \cdots + X_n)/n$ and define

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

then $Z_n \overset{D}{\to} Z$ where $Z \sim N(0,1)$. The sequence Z_n converges in distribution to a random variable Z that follows a standard normal distribution, which means that

$$\lim_{n\to\infty} P(Z_n \leq z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

1.1.1 Interval estimates using the CLT

We can use this to form random intervals that contain μ with some desired probability. For a standard normal random variable Z, we can easily find ℓ and u so that $P(Z < \ell) = \alpha_\ell$ and $P(Z > u) = \alpha_u$ using the normal quantile function in `R`. In the following code, I call ℓ by `lower_Z` and u by `upper_Z` so that we can identify these lower and upper bounds as bounds for Z. Similarly, I will call α_ℓ by `alpha_Z_lower` and α_u by `alpha_Z_upper`.

```
alpha_Z_lower = 0.01
alpha_Z_upper = 0.03
lower_Z = qnorm(alpha_Z_lower)
upper_Z = qnorm(alpha_Z_upper)
```

If we let $\alpha = \alpha_\ell + \alpha_u$, then we know that $P(\ell < Z < u) = 1 - \alpha$. We can use this to approximate the same probability for Z_n when n is large by invoking the CLT

$$P(\ell < Z_n < u) \approx P(\ell < Z < u) = 1 - \alpha.$$

The quantity Z_n is a function of both the data and the unknown population mean μ. We can rearrange the inequalities in $\ell < Z_n < u$ to isolate μ and get two an interval that will contain the true value μ with approximate probability $1 - \alpha$.

$$\begin{aligned} \ell < Z_n < u &= \ell < \tfrac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < u \\ &= \tfrac{\sigma}{\sqrt{n}}\ell < \overline{X}_n - \mu < \tfrac{\sigma}{\sqrt{n}}u \\ &= \tfrac{\sigma}{\sqrt{n}}\ell - \overline{X}_n < -\mu < \tfrac{\sigma}{\sqrt{n}}u - \overline{X}_n \\ &= -\tfrac{\sigma}{\sqrt{n}}\ell + \overline{X}_n > \mu > -\tfrac{\sigma}{\sqrt{n}}u + \overline{X}_n \end{aligned}$$

We get the random interval

$$(L, U) = \left(\overline{X}_n - \frac{\sigma}{\sqrt{n}}u, \overline{X}_n - \frac{\sigma}{\sqrt{n}}\ell\right)$$

which has approximate probability $1 - \alpha$ of containing μ. Using the symmetry of the standard normal, it is easy to see that this interval leave α_u below its lower bound and α_ℓ above its upper bound. This "change" of what is above and what is below is because the formula for Z_n has $-\mu$ in it and so we had to divide by -1 when isolating μ using the inequalities.

Let's rename things a bit. We will reference u and ℓ as z-scores. Let z_p be the z-score with p probability below it, so $P(Z \leq z_p) = p$. Let α_L be the probabity we want below our interval estimate and $\alpha_U$be the probability that we want above our interval estimate. The the interval we have is

$$(L, U) = \left(\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{1-\alpha_L}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{1-\alpha_U}\right)$$

and it looks more transparent in terms of what different pieces mean, at least to me. The *coverge* of the interval is $1 - \alpha$ where $\alpha = \alpha_L + \alpha_U$.

1.1.2 Setting sample size

If we set $\alpha_L = \alpha_U = \alpha/2$, then we get the symmetric interval

$$(L, U) = \left(\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right)$$

and its length is $\frac{2\sigma}{\sqrt{n}}z_{1-\alpha/2}$. If we want to get a symmetric interval with length at most δ, then we can set

$$\delta \geq \frac{2\sigma}{\sqrt{n}}z_{1-\alpha/2}$$

and isolate n get

$$n \geq \frac{4\sigma^2}{\delta^2}z^2_{1-\alpha/2}.$$

Any integer n satisfying this inequality will provide for us an interval that is smaller than δ. The smallest such n is

$$n_0 = \left\lceil \frac{4\sigma^2}{\delta^2}z^2_{1-\alpha/2} \right\rceil$$

where $\lceil x \rceil$ is the ceiling function.

1.1.3 Rejecting values of μ

We can make a test using our interval estimate. We reject the null hypothesis if it does not overlap with our interval estimate.

If we want to perform the hypothesis test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ and level α_0, then we can set $\alpha_L = \alpha_0$ and $\alpha_U = 0$. Our interval is of the form (L, ∞). The test is then to reject H_0 if $\mu_0 \leq L$. That is, if the sample mean $\overline{x}_n$ is too big, then we reject H_0.

If we want to perform the hypothesis test $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$ and level α_0, then we can set $\alpha_L = 0$ and $\alpha_U = \alpha_0$. Our interval is of the form $(-\infty, U)$. The test is then to reject H_0 if $\mu_0 \geq U$. That is, if the sample mean $\overline{x}_n$ is too small, then we reject H_0.

If we want to perform the hypothesis test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ and level α_0, then we can set $\alpha_L = \alpha_0/2$ and $\alpha_U = \alpha_0/2$. Our interval is of the form (L, U). The test is then to reject H_0 if $\mu_0 \leq L$ or $\mu_0 \geq U$. That is, if the sample mean $\overline{x}_n$ is too small or too large, then we reject H_0.

1.1.4 Example: Student's-t random variables

If the X_i are iid Student's-t(μ, s, ν) random variables with degrees of freedom $\nu > 2$ and scaling $s > 0$ that are known and center μ that is unknown. We have $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \frac{s^2\nu}{\nu-2}$. So the variance is known and we can set $\sigma^2 = \frac{s^2\nu}{\nu-2}$ and use the formulas from above. In R we can run the following code to the above tasks.

```
#setting seed
set.seed(1234567890)

# given things
n = 78
nu = 7
s = 3
mu = 5 # just for generation of data, we will not treat this as known
x = rt(n,df=nu)*s + mu

# left interval
alpha = 0.01
sigma = s*sqrt(nu/(nu-2))
left_int = c(-Inf,mean(x)+qnorm(1-alpha)*sigma/sqrt(length(x)))
print(left_int)
```

```
## [1]     -Inf 6.400291
```

```
left_int_rejected_mu_values = c(mean(x)+qnorm(1-alpha)*
                                  sigma/sqrt(length(x)),Inf)
print(left_int_rejected_mu_values)
```

```
## [1] 6.400291      Inf
```

```
# right interval
alpha = 0.01
sigma = s*sqrt(nu/(nu-2))
right_int = c(mean(x)-qnorm(1-alpha)*sigma/sqrt(length(x)),Inf)
print(right_int)
```

```
## [1] 4.530286      Inf
```

```
right_int_rejected_mu_values = c(-Inf,mean(x)-qnorm(1-alpha)*
                                   sigma/sqrt(length(x)))
print(right_int_rejected_mu_values)
```

```
## [1]     -Inf 4.530286
```

```
# symmetric interval
alpha = 0.01
sigma = s*sqrt(nu/(nu-2))
symmetric_int = c(mean(x)-qnorm(1-alpha/2)*sigma/sqrt(length(x)),
                  mean(x)+qnorm(1-alpha/2)*sigma/sqrt(length(x)))
print(symmetric_int)
```

```
## [1] 4.430015 6.500562
```

```
symmetric_int_rejected_mu_values_left = c(-Inf,mean(x)-qnorm(1-alpha/2)*
                                            sigma/sqrt(length(x)))
print(symmetric_int_rejected_mu_values_left)
```

```
## [1]     -Inf 4.430015
```

```
symmetric_int_rejected_mu_values_right = c(mean(x)+qnorm(1-alpha/2)*
                                             sigma/sqrt(length(x)),Inf)
print(symmetric_int_rejected_mu_values_right)
```

```
## [1] 6.500562      Inf
```

```
# Sample size for symmetric interval of length 2
alpha = 0.01
delta = 2
sigma = s*sqrt(nu/(nu-2))
n_0 = ceiling((2*sigma*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 84
```

```
# Sample size for symmetric interval of length 1
alpha = 0.01
delta = 1
sigma = s*sqrt(nu/(nu-2))
n_0 = ceiling((2*sigma*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 335
```

```
# Sample size for symmetric interval of length 0.5
alpha = 0.01
delta = 0.5
sigma = s*sqrt(nu/(nu-2))
n_0 = ceiling((2*sigma*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 1338
```

1.2 CLT using Slutsky's Theorem

We can use the CLT in the exact same way we did before, just with using an estimate of the variance. Let $\widehat{\sigma}_n^2$ be any consistent sequence of estimators for the population variance ($\widehat{\sigma}_n^2 \to \sigma^2$ in probability). Define Z_n by

$$Z_n = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\widehat{\sigma}}.$$

Slutsky's Theorem says that $Z_n \to Z$ in distribution where $Z \sim N(0,1)$.

The estimate of the variance can come through two processes. First, we can use an estimator using computed statistics if we know the functional form of the variance. Second, we can just use some generic estimator like the plug-in estimator.

1.2.1 Using functional form of the variance

This is best explored using an example. We will us the Exponential distribution with mean λ. The variance is λ^2. Our plug-in estimator of λ is $\overline{x}_n$, and so a consistent estimator of the variance is $\overline{x}_n^2$. Let's repeat the same `R` code as before for this example. We will keep using μ to represent the population mean, and so $\mu = \lambda$.

```
#setting seed
set.seed(1234567890)

# given things
n = 145
lambda = 5 # just for generation of data, we will not treat this as known
# remember that mu = lambda
x = rexp(n,rate=1)*lambda

# left interval
alpha = 0.1
sigma_hat = mean(x)
left_int = c(0,mean(x)+qnorm(1-alpha)*sigma_hat/sqrt(length(x)))
print(left_int)
```

```
## [1] 0.000000 5.247453
```

```
left_int_rejected_mu_values = c(mean(x)+qnorm(1-alpha)*
                                  sigma_hat/sqrt(length(x)),Inf)
print(left_int_rejected_mu_values)
```

```
## [1] 5.247453      Inf
```

```
# right interval
alpha = 0.1
sigma_hat = mean(x)
right_int = c(mean(x)-qnorm(1-alpha)*sigma_hat/sqrt(length(x)),Inf)
print(right_int)
```

```
## [1] 4.237949      Inf
```

```
right_int_rejected_mu_values = c(0,mean(x)-qnorm(1-alpha)*
                                   sigma_hat/sqrt(length(x)))
print(right_int_rejected_mu_values)
```

```
## [1] 0.000000 4.237949
```

```
# symmetric interval
alpha = 0.1
sigma_hat = mean(x)
symmetric_int = c(mean(x)-qnorm(1-alpha/2)*sigma_hat/sqrt(length(x)),
                  mean(x)+qnorm(1-alpha/2)*sigma_hat/sqrt(length(x)))
print(symmetric_int)
```

```
## [1] 4.094859 5.390543
```

```
symmetric_int_rejected_mu_values_left = c(0,mean(x)-qnorm(1-alpha/2)*
                                            sigma_hat/sqrt(length(x)))
print(symmetric_int_rejected_mu_values_left)
```

```
## [1] 0.000000 4.094859
```

```
symmetric_int_rejected_mu_values_right = c(mean(x)+qnorm(1-alpha/2)*
                                             sigma_hat/sqrt(length(x)),Inf)
print(symmetric_int_rejected_mu_values_right)
```

```
## [1] 5.390543      Inf
```

```
# Sample size for symmetric interval of length 2
alpha = 0.1
delta = 2
sigma_hat = mean(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 61
```

```
# Sample size for symmetric interval of length 1
alpha = 0.1
delta = 1
sigma_hat = mean(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 244
```

```
# Sample size for symmetric interval of length 0.5
alpha = 0.1
delta = 0.5
sigma_hat = mean(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 974
```

1.2.2 Using a plug-in estimator of the variance

In this scenario, we do not know the functional form of the variance, but we do know that it is finite. The best we can do is just estimate the variance using the data with an estimator like

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

We can get a consistent sequence of estimators using `sigma_hat = sd(x)` in `R` (the division is by $n-1$ instead of n, but who cares?). Let's go through all of the previous code with a real data set whose distribution we do not know, but we assume is of finite variance.

```
data("discoveries")
x = as.numeric(discoveries)

# left interval
alpha = 0.04
sigma_hat = sd(x)
left_int = c(0,mean(x)+qnorm(1-alpha)*sigma_hat/sqrt(length(x)))
print(left_int)
```

```
## [1] 0.000000 3.494616
```

```
left_int_rejected_mu_values = c(mean(x)+qnorm(1-alpha)*
                                  sigma_hat/sqrt(length(x)),Inf)
print(left_int_rejected_mu_values)
```

```
## [1] 3.494616      Inf
```

```
# right interval
alpha = 0.04
sigma_hat = sd(x)
right_int = c(mean(x)-qnorm(1-alpha)*sigma_hat/sqrt(length(x)),Inf)
print(right_int)
```

```
## [1] 2.705384      Inf
```

```
right_int_rejected_mu_values = c(0,mean(x)-qnorm(1-alpha)*
                                   sigma_hat/sqrt(length(x)))
print(right_int_rejected_mu_values)
```

```
## [1] 0.000000 2.705384
```

```
# symmetric interval
alpha = 0.04
sigma_hat = sd(x)
symmetric_int = c(mean(x)-qnorm(1-alpha/2)*sigma_hat/sqrt(length(x)),
                  mean(x)+qnorm(1-alpha/2)*sigma_hat/sqrt(length(x)))
print(symmetric_int)
```

```
## [1] 2.637072 3.562928
```

```
symmetric_int_rejected_mu_values_left = c(0,mean(x)-qnorm(1-alpha/2)*
                                            sigma_hat/sqrt(length(x)))
print(symmetric_int_rejected_mu_values_left)
```

```
## [1] 0.000000 2.637072
```

```
symmetric_int_rejected_mu_values_right = c(mean(x)+qnorm(1-alpha/2)*
                                             sigma_hat/sqrt(length(x)),Inf)
print(symmetric_int_rejected_mu_values_right)
```

```
## [1] 3.562928      Inf
```

```
# Sample size for symmetric interval of length 1
alpha = 0.04
delta = 1
sigma_hat = sd(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 86
```

```
# Sample size for symmetric interval of length 0.5
alpha = 0.04
delta = 0.5
sigma_hat = sd(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 343
```

```
# Sample size for symmetric interval of length 0.1
alpha = 0.04
delta = 0.1
sigma_hat = sd(x)
n_0 = ceiling((2*sigma_hat*qnorm(1-alpha/2)/delta)^2)
print(n_0)
```

```
## [1] 8573
```

1.3 Wald Tests

A Wald test uses a test statistic of the form

$$Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\text{Var}\left(\widehat{\theta}_n\right)}}$$

and in our case we have always used $\theta = \mu$, the population mean, and $\widehat{\theta} = \overline{x}_n$, the sample mean. When the x_i are independent and identically distributed and have common variance σ^2, then the denominator is $\sigma/\sqrt{n}$. Just as before, we can invoke Slutsky's theorem and use a consistent estimator of the variance $\widehat{\sigma}^2$. All error rates are calculated relative to the asymptotic normality of the test statistic Z_n.

1.3.1 Tests of inequality

When testing the null hypothesis $H_0 : \mu \leq \mu_0$ versus the alternative $H_1 : \mu > \mu_0$, the null hypothesis is rejected when Z_n is larger than some value Z_{cutoff}. The Type I Error Rate (TIER, α, significance level) of

the test is given by

$$\alpha = \max_{\mu \leq \mu_0} P(Z_n > Z_{\text{cutoff}} | \mu) = P(Z_n > Z_{\text{cutoff}} | \mu = \mu_0)$$

and so if we want to set α at some specified value, then we need to set Z_{cutoff} so that $\alpha = P(Z_n > Z_{\text{cutoff}} | \mu = mu_0)$. To do this, we can use the asymptotic normality and set Z_{cutoff} by

$$Z_{\text{cutoff}} = z_{1-\alpha}$$

where z_{prob} is the prob quantile from a standard normal distribution. Formally, if $Z \sim N(0,1)$, then z_{prob} is defined by

$$\text{prob} = P(Z \leq z_{\text{prob}}).$$

In terms of the statistics $T_n = \overline{x}_n$, the rejection region is the set

$$\text{Rejection Region } = \left(\mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, \infty \right)$$

where we can substitute in a consistent estimator for σ by invoking Slutsky's Theorem.

To get a confidence interval, we can invert the test. We reject H_0 if

$$\overline{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}},$$

which happens if

$$\overline{x}_n - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} > \mu_0.$$

The complement of this is

$$\overline{x}_n - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} < \mu_0$$

if we ignore the possibity of equality (which has probability 0 in the limit because of the CLT and the fact that a normal random variable is continuous). The confidence interval is given by

$$\text{Confidence Interval } = \left(\overline{x}_n - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, \infty \right).$$

For a test of the null hypothesis $H_0 : \mu \geq \mu_0$ versus the alternative $H_1 : \mu < \mu_0$, the rejection region and confidence interval are given by

$$\begin{aligned} \text{Rejection Region} &= \left(-\infty, \mu_0 - z_{1-\alpha} \tfrac{\sigma}{\sqrt{n}} \right) \\ \text{Confidence Interval} &= \left(-\infty, \overline{x}_n + z_{1-\alpha} \tfrac{\sigma}{\sqrt{n}} \right). \end{aligned}$$

1.3.2 Test of equality

When testing the null hypothesis $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu \neq \mu_0$, the null hypothesis is rejected when $|Z_n|$ is larger than some value Z_{cutoff}. Note that we are doing a symmetric test here, we could try to do something else, but this is usually the test done. The Type I Error Rate (TIER, α, significance level) of the test is given by

$$\alpha = P(|Z_n| > Z_{\text{cutoff}} | \mu = \mu_0) = P(Z_n > Z_{\text{cutoff}} | \mu = \mu_0) + P(Z_n < -Z_{\text{cutoff}} | \mu = \mu_0)$$

and so if we want to set α at some specified value, then we need to set Z_{cutoff} so that $\frac{\alpha}{2} = P(Z_n > Z_{\text{cutoff}} | \mu = mu_0)$. To do this, we can use the asymptotic normality and set Z_{cutoff} by

$$Z_{\text{cutoff}} = z_{1-\frac{\alpha}{2}}$$

where z_{prob} is the prob quantile from a standard normal distribution. Formally, if $Z \sim N(0,1)$, then z_{prob} is defined by

$$\text{prob} = P(Z \leq z_{\text{prob}}).$$

In terms of the statistics $T_n = \overline{x}_n$, the rejection region is the set

$$\text{Rejection Region } = \left(-\infty, \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) \cup \left(\mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \infty\right)$$

where we can substitute in a consistent estimator for σ by invoking Slutsky's Theorem.

To get a confidence interval, we can invert the test. We reject H_0 if

$$\begin{aligned} \overline{x}_n &> \mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\ &\text{or} \\ \overline{x}_n &< \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \end{aligned}$$

which happens if

$$\begin{aligned} \overline{x}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &> \mu_0 \\ &\text{or} \\ \overline{x}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &< \mu_0. \end{aligned}$$

The complement of this is

$$\begin{aligned} \overline{x}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &< \mu_0 \\ &\text{and} \\ \overline{x}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &> \mu_0. \end{aligned}$$

if we ignore the possibity of equality (which has probability 0 in the limit because of the CLT and the fact that a normal random variable is continuous). The confidence interval is given by

$$\text{Confidence Interval } = \left(\overline{x}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right),$$

which is often written in short-hand by

$$\text{Confidence Interval } = \overline{x}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

2 Known variance problems

Question 1

Suppose that the data are from a Logistic distribution with center μ and scaling $s > 0$. The density for a single datum

$$f_X(x) = \frac{1}{s\left(\exp(0.5(x-\mu)/s) + \exp(-0.5(x-\mu)/s)\right)^2}.$$

The expected value μ is unknown and the scaling is known to be $s = 3$. The variance of an individual draw from the population is $\sigma^2 = s^2\pi^2/3$. The data `x_q1` contains independent and identically distributed Logistic draws with scaling $s = 3$. Use these data and the CLT approximation to form an interval $(-\infty, U)$ that has 98% probability of containing the true value of μ.

Question 2

Suppose that the data are from a Logistic III distribution with center μ, scaling $s > 0$, and shape $\alpha > 0$. The density for a single datum

$$f_X(x) = \frac{\Gamma(2\alpha)}{s\Gamma(\alpha)^2 \left(\exp(0.5(x-\mu)/s) + \exp(-0.5(x-\mu)/s)\right)^{2\alpha}}.$$

The expected value μ is unknown, the scaling is known to be $s = 10$, and the shape is known to be $\alpha = 15$. The variance of an individual draw from the population is $\sigma^2 = 2s^2\psi_1(\alpha) \approx 0.138 \times s^2$ for $\alpha = 15$ and where ψ_1 is the trigamma function (the second derivative of the logarithm of the gamma funtion, oof, do not worry about it). The data `x_q2` contains independent and identically distributed Logistic III draws with scaling $s = 10$ and shape $\alpha = 15$. Use these data and the CLT approximation to form a symmetric interval (L, U) that has 99% probability of containing the true value of μ.

Question 3

Suppose that the data are from a Hyperbolic distribution with center μ, scaling $s > 0$, and shape parameter $\alpha > 0$. The density for a single datum

$$f_X(x) = \frac{1}{2sK_1(\alpha)} \exp\left(-\alpha\sqrt{1 + \frac{(x-\mu)^2}{s^2}}\right)$$

where K_1 is a modified Bessel function of the second kind (do not worry about it, it is a constant in our density). The expected value μ is unknown, the scaling is known to be $s = 2$, and shape is known to be $\alpha = 1$. The variance of an individual draw from the population is $\sigma^2 = \frac{s^2 K_2(\alpha)}{\alpha K_1(\alpha)} \approx 0.54 \times s^2$ for $\alpha = 1$. The data `x_q3` contains independent and identically distributed Hyperbolic draws with scaling $s = 2$ and shape $\alpha = 1$. Use these data and the CLT approximation to form an interval (L, ∞) that has 99% probability of containing the true value of μ.

Question 4

Suppose that the data are from a Laplace distribution with center μ and scaling $s > 0$. The density for a single datum

$$f_X(x) = \frac{1}{2s} \exp\left(-\frac{|x-\mu|}{s}\right).$$

The expected value μ is unknown and the scaling is known to be $s = 3$. The variance of an individual draw from the population is $\sigma^2 = 2s^2$. What is the minimal sample size needed so that the CLT approximation for the distribution of the sample mean can be used to form a symmetric interval (L, U) that has 99.5% probability of containing the true value of μ and has length at most 0.2?

Question 5

Suppose that the data are from a Student's-t distribution with center μ, scaling $s > 0$, and degress of freedom $\nu > 0$. The density for a single datum

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu s^2}} \left(1 + \frac{(x-\mu)^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}}.$$

The expected value μ exists when $\nu > 1$ and the variance exists when $\nu > 2$. Assume that ν is known to be $\nu = 5$ and the scaling is known to be $s = 3$, but that the mean μ is unknown. The variance of an individual draw from the population is $\sigma^2 = \frac{\nu}{\nu-2}s^2$ when it exists. What is the rejection region for $\overline{x}_n$ when performing a Wald test of $H_0 : \mu \leq 12$ versus $H_1 : \mu > 12$ with significance level $\alpha = 0.02$ and $n = 35$ using the CLT approximation?

3 Variance as a function of the mean problems

Question 6

Suppose that the data are from a Negative Binomial distribution with mean $\mu > 0$ and size $r > 0$. The mass for a single datum is

$$f_X(x) = \binom{x + r - 1}{x} \frac{r^r \mu^x}{(r + \mu)^{r+x}}$$

for non-negative integer x. The expected value μ is unknown and the size is known to be $r = 10$. The variance of an individual draw from the population is $\sigma^2 = \mu(1 + \mu/r) = \mu(1 + \mu/10)$ for $r = 10$. The data `x_q6` contains independent and identically distributed Negative Binomial draws with size $r = 10$. Use these data and the CLT approximation to form an interval (L, ∞) that has 99% probability of containing the true value of μ. Make sure to use the functional form of the variance with a plug-in estimator for μ.

Question 7

Suppose that the data are from a Rayleigh distribution with mean μ. The density for a single datum is

$$f_X(x) = \frac{\pi x}{2\mu^2} \exp\left(-\frac{\pi x^2}{4\mu^2}\right)$$

for $x > 0$. The expected value μ is unknown. The variance of an individual draw from the population is $\sigma^2 = \frac{4-\pi}{\pi}\mu^2$. The data `x_q7` contains independent and identically distributed Rayleigh draws. Use these data and the CLT approximation to form an interval $(-\infty, U)$ that has 94% probability of containing the true value of μ. Make sure to use the functional form of the variance with a plug-in estimator for μ.

Question 8

Suppose that the data are from a Inverse-Gaussian distribution with mean $\mu > 0$ and shape $\lambda > 0$. The density for a single datum is

$$f_X(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x - \mu)^2}{2\mu^2 x}\right)$$

for $x > 0$. The expected value μ is unknown and the shape is known to be $\lambda = 2.5$. The variance of an individual draw from the population is $\sigma^2 = \frac{\mu^3}{\lambda}$. The data `x_q8` contains independent and identically distributed Inverse-Gaussian draws with shape $\lambda = 2.5$. Use these data and the CLT approximation to form an interval (L, U) that has 97% probability of containing the true value of μ. Make sure to use the functional form of the variance with a plug-in estimator for μ.

Question 9

Suppose that the data are from a Lomax distribution with mean $\mu > 0$ and shape $\alpha > 0$. The density for a single datum is

$$f_X(x) = \frac{\alpha\lambda^\alpha}{(\alpha-1)(\mu+(\alpha-1)x)^{\alpha+1}}$$

for $x > 0$. The expected value is finite when $\alpha > 1$ and the variance is finite if $\alpha > 2$. The expected value μ is unknown and the shape is known to be $\alpha = 7$. The variance of an individual draw from the population is $\sigma^2 = \frac{\alpha}{\alpha-2}\mu^2$. What is the minimal sample size needed so that the CLT approximation for the distribution of the sample mean can be used to form a symmetric interval (L, U) that has 97.5% probability of containing the true value of μ and has length at most 0.1 when we think a reasonable guess for the mean (based on previous experiments) is $\widehat{\mu} = 5$? Make sure to use the functional form of the variance using the estimate for μ from the previous experiment.

Question 10

Suppose that the data are from a Poisson distribution with mean $\mu > 0$. The mass for a single datum is

$$f_X(x) = \frac{\mu^x}{x!}\exp(-\mu)$$

for non-negative integer x. The mean μ is unknown. The variance of an individual draw from the population is $\sigma^2 = \mu$ when it exists. What is the rejection region for $\overline{x}_n$ when performing a Wald test of $H_0 : \mu = 7.5$ versus $H_1 : \mu \neq 7.5$ with significance level $\alpha = 0.08$ and $n = 102$ using the CLT approximation? Make sure to use the functional form of the variance as a function of μ and use that fact that the rejection region is formed when it is assumed that the null hypothesis is true.

4 Unknown and estimated variance problems

Question 11

You have collected data of size $n = 48$ that has sample mean $\overline{x} = 4.68$ and sample variance $\hat{\sigma}^2 = 2.35$. Form a 96% confidence interval of the form (L, ∞) for the population mean.

Question 12

You have collected the data that come from the commands

```
library(boot) #you should have this library installed
data(gravity)
x_q11 = gravity$g
```

Form a symmetric 99% confidence interval of the form (L, U) for the population mean.

Question 13

You have collected the data that come from the commands

```
library(boot) #you should have this library installed
data(neuro)
x_q12 = neuro[,3]
```

Form a 96% confidence interval of the form $(-\infty, U)$ for the population mean.

Question 14

You are planning to collect a data set of size n. Previous experiments suggest that a very good estimate for the population variance is $\hat{\sigma}^2 = 6.8$. How many samples will be needed for a symmetric 95% confidence interval to have length at most 0.2?

Question 15

You are planning to collect a data set of size $n = 72$. Previous experiments suggest that a very good estimate for the population variance is $\hat{\sigma}^2 = 3.7$. What is the rejection region for $\overline{x}_n$ when performing a Wald test of $H_0 : \mu \geq 4.5$ versus $H_1 : \mu < 4.5$ with significance level $\alpha = 0.02$ using the CLT approximation?