8.34. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson population distribution with mean $\lambda > 0$, where λ is unknown. We know $\overline{Y}$ is an unbiased estimator of λ. In other words,

$$E(\overline{Y}) = \lambda.$$

Therefore, $\overline{Y}$ is certainly a sensible estimator on the basis that it is unbiased. Furthermore, calculating the standard error of $\overline{Y}$ is easy. We have

$$V(\overline{Y}) = \frac{\lambda}{n} \implies \sigma_{\overline{Y}} = \sqrt{\frac{\lambda}{n}}.$$

Now, as is frequently the case, the standard error of our point estimator (here, $\overline{Y}$) depends on λ, which is an unknown population parameter. Therefore, we can *estimate* the standard error of $\overline{Y}$ by using

$$\widehat{\sigma}_{\overline{Y}} = \sqrt{\frac{\overline{Y}}{n}}.$$

Note that all we have done here is replace λ in the standard error with an unbiased estimator of it. Interestingly, because S^2 is also an unbiased estimator of λ, we could also estimate the standard error by using

$$\widehat{\sigma}_{\overline{Y}} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}}.$$

Either answer would be a reasonable way to estimate the standard error of $\overline{Y}$.

8.36. This problem is similar to Problem 8.34. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential population distribution with mean $\theta > 0$, where θ is unknown. We know $\overline{Y}$ is an unbiased estimator of θ. The standard error of $\overline{Y}$ is calculated as follows:

$$V(\overline{Y}) = \frac{\theta^2}{n} \implies \sigma_{\overline{Y}} = \sqrt{\frac{\theta^2}{n}} = \frac{\theta}{\sqrt{n}}.$$

Again, the standard error of our point estimator (here, $\overline{Y}$) depends on θ, which is an unknown population parameter. We can *estimate* the standard error of $\overline{Y}$ by using

$$\widehat{\sigma}_{\overline{Y}} = \frac{\overline{Y}}{\sqrt{n}}$$

Note that all we have done here is replace θ in the standard error with an unbiased estimator of it.

8.60. In this problem, we envision an iid sample of $n = 130$ "healthy" humans, where, on each individual, we measure

$$Y = \text{body temperature (measured in deg F)}.$$

(a) From the sample, we are given $\overline{y} = 98.25$ deg F and $s = 0.73$ deg F. The population distribution of body temperatures is not known, so we can use a large-sample interval for the population mean μ. A large-sample 99% confidence interval is

$$\overline{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \longrightarrow 98.25 \pm 2.58\left(\frac{0.73}{\sqrt{130}}\right) \longrightarrow (98.08, 98.42).$$

```
> qnorm(0.995,0,1) # upper 0.005 quantile from N(0,1)
[1] 2.575829
```

Interpretation: We are (approximately) 99% confident the population mean human body temperature μ is between 98.08 and 98.42 deg F.

(b) The confidence interval for μ does not contain 98.6 deg F, the "accepted" average temperature. What conclusions can we draw? This is hard to answer as there could be many explanations:

- Our inference procedure only utilizes 99% confidence. Therefore, the population mean could be $\mu = 98.6$ deg F, and this is one of the few intervals that would exclude it.
- It could be the population mean μ really is 98.6 deg F; it's just that our "sample" was not a random sample from the population; e.g., perhaps some of the individuals were not really "healthy."
- It could be the sample was representative and the population mean μ is slightly less than 98.6 deg F. The interval certainly does not provide evidence that μ is larger than 98.6 deg F.

8.65. In this problem, we envision two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a Bernoulli(p_1) population
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a Bernoulli(p_2) population.

Here, p_1 (p_2) is the population proportion of defective items from Line A (Line B). In part (a), our goal is to estimate the parameter $\theta = p_1 - p_2$, the difference of the population proportions. The problem gives the sample sizes $n_1 = n_2 = 100$ and the sample proportions (i.e., the point estimates); these are

$$\widehat{p}_1 = \frac{18}{100} = 0.18 \quad \text{and} \quad \widehat{p}_2 = \frac{12}{100} = 0.12.$$

We can use this information to write a large-sample (approximate) confidence interval for $\theta = p_1 - p_2$. We will use the interval

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

We have

$$(0.18-0.12) \pm 2.33\sqrt{\frac{0.18(1-0.18)}{100} + \frac{0.12(1-0.12)}{100}} \quad \longrightarrow \quad 0.06 \pm 0.117 \quad \longrightarrow \quad (-0.057, 0.177).$$

Interpretation: We are (approximately) 98% confident the difference of the population proportions $\theta = p_1 - p_2$ is between -0.057 and 0.177.

```
> qnorm(0.99,0,1) # upper 0.01 quantile from N(0,1)
[1] 2.326348
```

(b) Note that the interval does include "0." In other words, on the basis of this analysis, "0" is a plausible value for $\theta = p_1 - p_2$ as it resides in the confidence interval. Of course, if this is true (i.e., if $\theta = p_1 - p_2 = 0$), then the population proportions p_1 and p_2 would be equal.

8.87. In this problem, we envision two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ population distribution,

The goal is to estimate the parameter $\theta = \mu_1 - \mu_2$, the difference of the population mean prices in tuna packed with water (population 1) and tuna packed with oil (population 2). We have samples of size $n_1 = 14$ and $n_2 = 11$ from these populations.

In part (a), we are being asked to write a 90% confidence interval for $\theta = \mu_1 - \mu_2$. This can be done entirely in R, and we can request the equal-variance/unequal-variance intervals:

```
> water = c(0.99,1.92,1.23,0.85,0.65,0.69,0.60,0.53,1.41,1.12,0.63,0.67,0.60,0.66)
> oil = c(2.56,1.92,1.30,1.79,1.23,0.62,0.66,0.62,0.65,0.60,0.67)

> t.test(water,oil,conf.level=0.90,var.equal=TRUE)$conf.int
[1] -0.6229708  0.1212825

> t.test(water,oil,conf.level=0.90,var.equal=FALSE)$conf.int
[1] -0.6548617  0.1531734
```

There are minor differences in the intervals, but the overall message is the same. Note that the interval for $\theta = \mu_1 - \mu_2$ does include "0." In other words, on the basis of this analysis, "0" is a plausible value for $\theta = \mu_1 - \mu_2$ as it resides in the confidence interval. Of course, if this is true (i.e., if $\theta = \mu_1 - \mu_2 = 0$), then the population mean prices μ_1 and μ_2 would be equal.

8.95. In this problem, the random variable

$$Y = \text{noise level (measured in decibels)}$$

is measured on each six heavy trucks. We envision $Y_1, Y_2, ..., Y_6$ as an iid sample of size $n = 6$ from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, where both μ and σ^2 are unknown. Our goal is to write a 90% confidence interval for the population variance σ^2 on the basis of this sample. We will use the interval we derived in class; i.e.,

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right).$$

I coded the calculation of this interval in R:

```
> noise = c(85.4,86.8,86.1,85.3,84.8,86.0)
> ci.lower = 5*var(noise)/qchisq(0.95,5)
> ci.upper = 5*var(noise)/qchisq(0.05,5)
> round(c(ci.lower,ci.upper),1)
[1] 0.2 2.2
```

Interpretation: We are 90% confident the population variance σ^2 is between 0.2 and 2.2 $(\text{decibels})^2$.

8.127. In this problem, $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma(c_0, β) population distribution, where the shape parameter $\alpha = c_0$ is known and the scale parameter $\beta > 0$ is unknown. Our goal is to derive a confidence interval for β. The problem says "approximate" confidence interval, so this should trigger in your mind that the Central Limit Theorem will be used. Recall that in this population

$$\begin{aligned} \mu &= c_0\beta \\ \sigma^2 &= c_0\beta^2. \end{aligned}$$

Therefore, applying the CLT directly, the approximate sampling distribution of $\overline{Y}$ is

$$\overline{Y} \sim \mathcal{N}\left(c_0\beta, \ \frac{c_0\beta^2}{n}\right) \quad \Longrightarrow \quad Q = \frac{\overline{Y} - c_0\beta}{\sqrt{\frac{c_0\beta^2}{n}}} \sim \mathcal{AN}(0,1),$$

when the sample size n is large (e.g., like $n = 100$). Note that because (the large-sample) distribution of Q does not depend on any unknown population parameters, Q is a large-sample pivot. Therefore, we can write

$$\begin{aligned} 1-\alpha \approx P\left(-z_{\alpha/2} < \frac{\overline{Y} - c_0\beta}{\sqrt{\frac{c_0\beta^2}{n}}} < z_{\alpha/2}\right) &= P\left(-z_{\alpha/2} < \frac{\overline{Y} - c_0\beta}{\beta\sqrt{\frac{c_0}{n}}} < z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2} < \frac{\overline{Y}}{\beta\sqrt{\frac{c_0}{n}}} - \frac{c_0\beta}{\beta\sqrt{\frac{c_0}{n}}} < z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2} < \frac{\overline{Y}}{\beta\sqrt{\frac{c_0}{n}}} - \sqrt{c_0 n} < z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2} + \sqrt{c_0 n} < \frac{\overline{Y}}{\beta\sqrt{\frac{c_0}{n}}} < z_{\alpha/2} + \sqrt{c_0 n}\right) \\ &= P\left(\frac{\sqrt{\frac{c_0}{n}}\left(-z_{\alpha/2} + \sqrt{c_0 n}\right)}{\overline{Y}} < \frac{1}{\beta} < \frac{\sqrt{\frac{c_0}{n}}\left(z_{\alpha/2} + \sqrt{c_0 n}\right)}{\overline{Y}}\right) \\ &= P\left(\frac{\overline{Y}}{\sqrt{\frac{c_0}{n}}\left(-z_{\alpha/2} + \sqrt{c_0 n}\right)} > \beta > \frac{\overline{Y}}{\sqrt{\frac{c_0}{n}}\left(z_{\alpha/2} + \sqrt{c_0 n}\right)}\right) \\ &= P\left(\frac{\overline{Y}}{\sqrt{\frac{c_0}{n}}\left(z_{\alpha/2} + \sqrt{c_0 n}\right)} < \beta < \frac{\overline{Y}}{\sqrt{\frac{c_0}{n}}\left(-z_{\alpha/2} + \sqrt{c_0 n}\right)}\right). \end{aligned}$$

This argument shows that

$$\left(\frac{\overline{Y}}{\sqrt{\frac{c_0}{n}\left(z_{\alpha/2}+\sqrt{c_0 n}\right)}}, \frac{\overline{Y}}{\sqrt{\frac{c_0}{n}\left(-z_{\alpha/2}+\sqrt{c_0 n}\right)}} \right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for β. When $n = 100$, the lower endpoint is

$$\frac{\overline{Y}}{\sqrt{\frac{c_0}{100}\left(z_{\alpha/2}+\sqrt{100c_0}\right)}} = \frac{\overline{Y}}{\frac{\sqrt{c_0}z_{\alpha/2}}{10}+c_0} = \frac{\overline{Y}}{c_0+0.1z_{\alpha/2}\sqrt{c_0}},$$

as stated. The upper endpoint

$$\frac{\overline{Y}}{c_0-0.1z_{\alpha/2}\sqrt{c_0}}$$

is found similarly.

8.128. This problem deals with comparing two population means μ_1 and μ_2, from normal distributions, where the population variances obey $\sigma_2^2 = k\sigma_1^2$, where k is a known constant. Of course, if $k = 1$, then this is our "equal-variance" case. Specifically, suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, k\sigma_1^2)$ population distribution,

where all population parameters are unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, the difference of the two population means. The derivation will closely mirror what we did in the notes.

(a) We know

$$\overline{Y}_{1+} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_2, \frac{k\sigma_1^2}{n_2}\right).$$

Because $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$ are both normally distributed, the difference $\overline{Y}_{1+} - \overline{Y}_{2+}$ is too (i.e., the difference is a simple linear combination). Therefore, because the two samples are independent,

$$\overline{Y}_{1+} - \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{k\sigma_1^2}{n_2}\right).$$

Standardizing, we get

$$Z^* = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{k\sigma_1^2}{n_2}}} = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma_1\sqrt{\frac{1}{n_1} + \frac{k}{n_2}}} \sim \mathcal{N}(0,1).$$

(b) We also know

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{and} \quad \frac{(n_2-1)S_2^2}{k\sigma_1^2} \sim \chi^2(n_2-1).$$

Therefore, because the two samples are independent,

$$W^* = \frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{k\sigma_1^2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2/k}{\sigma_1^2} \sim \chi^2(n_1+n_2-2).$$

(c) Because $Z^* \perp\!\!\!\perp W^*$ (why?), we have

$$T^* = \frac{\dfrac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1-\mu_2)}{\sigma_1\sqrt{\dfrac{1}{n_1}+\dfrac{k}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2/k}{\sigma_1^2}} \Big/ (n_1+n_2-2)} \sim t(n_1+n_2-2).$$

However, note that we can write T^* above as

$$T^* = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1-\mu_2)}{S_p^*\sqrt{\dfrac{1}{n_1}+\dfrac{k}{n_2}}},$$

where

$$S_p^* = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2/k}{n_1+n_2-2}}.$$

(d) Pivoting off T^*, we can write

$$1-\alpha = P\left(-t_{n_1+n_2-2,\alpha/2} < \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1-\mu_2)}{S_p^*\sqrt{\dfrac{1}{n_1}+\dfrac{k}{n_2}}} < t_{n_1+n_2-2,\alpha/2} \right).$$

After performing the usual algebra; i.e., to isolate $\mu_1-\mu_2$ in the center of the inequality, we conclude

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} S_p^* \sqrt{\frac{1}{n_1}+\frac{k}{n_2}}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$.

(e) If $k=1$, then this is our "equal-variance" case, which was covered in the notes.

8.132. In this problem, $Y_1, Y_2, ..., Y_n$ is an iid sample from a power family population distribution with cumulative distribution function

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \left(\dfrac{y}{\theta}\right)^\alpha, & 0 \le y \le \theta \\ 1, & y > \theta. \end{cases}$$

In this population model, α is assumed to be known (with $\alpha = c$) and $\theta > 0$ is unknown. Our goal is to derive a $100(1-\alpha)\%$ confidence interval for θ.

(a) For $0 \leq y \leq \theta$, the cdf of the maximum order statistic $Y_{(n)}$ is

$$\begin{aligned} F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) &= P(Y_1 \leq y, Y_2 \leq y, ..., Y_n \leq y) \\ &= P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) = [F_Y(y)]^n = \left[\left(\frac{y}{\theta}\right)^c\right]^n = \left(\frac{y}{\theta}\right)^{nc}. \end{aligned}$$

Summarizing,

$$F_{Y_{(n)}}(y) = \begin{cases} 0, & y < 0 \\ \left(\frac{y}{\theta}\right)^{nc}, & 0 \leq y \leq \theta \\ 1, & y > \theta. \end{cases}$$

(b) Define

$$Q = \frac{Y_{(n)}}{\theta}.$$

Note that

$$0 \leq y_{(n)} \leq \theta \iff q = \frac{y_{(n)}}{\theta} \in [0, 1].$$

Therefore, the support of $Q = Y_{(n)}/\theta$ is

$$R_Q = \{q : 0 \leq q \leq 1\}.$$

For $0 \leq q \leq 1$, the cdf of Q is

$$F_Q(q) = P(Q \leq q) = P\left(\frac{Y_{(n)}}{\theta} \leq q\right) = P(Y_{(n)} \leq \theta q) = F_{Y_{(n)}}(\theta q) = \left(\frac{\theta q}{\theta}\right)^{nc} = q^{nc}.$$

Summarizing,

$$F_Q(q) = \begin{cases} 0, & q < 0 \\ q^{nc}, & 0 \leq q \leq 1 \\ 1, & q > 1. \end{cases}$$

Because the distribution of Q (as described by its cdf) does not depend on any unknown population parameters, Q is a pivotal quantity. Note further that

$$P(k < Q < 1) = P\left(k < \frac{Y_{(n)}}{\theta} < 1\right) = F_Q(1) - F_Q(k) = 1 - k^{nc}.$$

(c) With $n = 5$ and $c = 2.4$, we have

$$0.95 \stackrel{\text{set}}{=} P\left(k < \frac{Y_{(5)}}{\theta} < 1\right) = 1 - k^{12} \implies k^{12} = 0.05 \implies k = (0.05)^{1/12} \approx 0.779.$$

Therefore, we have

$$\begin{aligned} 0.95 = P\left(0.779 < \frac{Y_{(5)}}{\theta} < 1\right) &= P\left(\frac{1}{0.779} > \frac{\theta}{Y_{(5)}} > 1\right) \\ &= P\left(\frac{Y_{(5)}}{0.779} > \theta > Y_{(5)}\right) = P\left(Y_{(5)} < \theta < \frac{Y_{(5)}}{0.779}\right). \end{aligned}$$

This argument shows that

$$\left(Y_{(5)}, \frac{Y_{(5)}}{0.779}\right)$$

is a 95% confidence interval for θ.

8.134. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, where both parameters are unknown. We derived a $100(1-\alpha)\%$ confidence interval for μ by using the t distribution; in particular,

$$\left(\overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. The width of this interval (i.e., how long it is) is the upper endpoint minus the lower endpoint; i.e., the width W is

$$W = \left(\overline{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) - \left(\overline{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 2t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

We want to calculate the expected width; i.e., $E(W)$. Note that

$$E(W) = E\left(2t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = \frac{2t_{n-1,\alpha/2}}{\sqrt{n}}E(S),$$

where S is the sample standard deviation. In Problem 8.16 (HW6), we showed that S is a biased estimator of σ; in particular,

$$E(S) = \left[\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}\right]\sigma.$$

Therefore, the expected width of the t confidence interval for μ is

$$E(W) = \frac{2t_{n-1,\alpha/2}}{\sqrt{n}}E(S) = \frac{2t_{n-1,\alpha/2}}{\sqrt{n}}\left[\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}\right]\sigma = \left[\frac{2\sqrt{2}t_{n-1,\alpha/2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n(n-1)}\Gamma(\frac{n-1}{2})}\right]\sigma.$$

For example, if $n = 10$ and $\alpha = 0.05$ (i.e., a 95% confidence interval), then

$$E(W) = \left[\frac{2\sqrt{2}t_{9,0.025}\Gamma\left(5\right)}{\sqrt{90}\Gamma(4.5)}\right]\sigma \approx 1.39\sigma.$$

The constant above can be calculated in R:

```
> constant = 2*sqrt(2)*qt(0.975,9)*gamma(5)/(sqrt(90)*gamma(4.5))
> constant
[1] 1.391597
```