1. Suppose that you want to estimate the mean pH of rainfalls in an area that suffers from heavy pollution due to the discharge of smoke from a power plant. Assume that $\sigma$ is in the neighborhood of .5 pH and that you want the sample mean to lie within .02 of $\mu$ with probability near .95. Approximately how many rainfalls must be included in your sample (one pH reading per rainfall)? (SET UP ONLY) (12 pts)

A $(1 - \alpha)100\%$ CI for $\mu$ is

$$Y \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$ 

We want

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = B$$ 

where

- $\alpha = .05 \implies z_{\alpha/2} = z_{.025} = 1.96$
- $\sigma = .5$
- $B = .02$

Thus

$$\frac{(1.96)(.5)}{\sqrt{n}} = .02 \implies \sqrt{n} = \frac{(1.96)(.5)}{.02} \implies n = \left(\frac{(1.96)(.5)}{.02}\right)^2$$
2. A Swedish study considered the effect of low-dose aspirin on reducing the risk of stroke and heart attacks among people who have already suffered a stroke. Of 1360 patients, 676 were randomly assigned to the aspirin treatment and 684 to a placebo treatment. During a follow up period of three years, the number of death due to myocardial infarction were 18 for the aspirin group and 28 for the placebo group. Let \( p_1 \) denote the proportion of all patients who would die from myocardial infarction within three years if under the aspirin regime, and let \( p_2 \) denote the proportion of all patients who would die from myocardial infarction within three years if under the placebo regime. Calculate a 92% confidence interval for \( p_1 - p_2 \). 

\[
\hat{p}_1 = \frac{18}{676} \quad \hat{p}_2 = \frac{28}{684} \\
\hat{q}_1 = \frac{658}{676} \quad \hat{q}_2 = \frac{656}{684}
\]

\( \alpha = 0.08 \implies z_{\alpha/2} = z_{0.04} = 1.75 \)

Thus a 98% CI for \( p_1 - p_2 \) is

\[
(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = \left( \frac{18}{676} - \frac{28}{684} \right) \pm (1.75) \sqrt{\left( \frac{18}{676} \right) \left( \frac{658}{676} \right) + \left( \frac{28}{684} \right) \left( \frac{656}{684} \right)}
\]
3. Let \( \hat{\theta} \) be an estimator of \( \theta \). Define the bias, \( B(\hat{\theta}) \), and the mean square error, \( \text{MSE}(\hat{\theta}) \), of \( \hat{\theta} \) and show that

\[
\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2.
\]

(12 pts)

\[
B(\hat{\theta}) = E(\hat{\theta}) - \theta
\]

\[
\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]
\]

An obvious but handy fact:

\[
\hat{\theta} - \theta = [\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta].
\]

Thus

\[
\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]
\]

by defn

\[
= E\left(\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2\right)
\]

by handy fact

\[
= E\left([\hat{\theta} - E(\hat{\theta})]^2 + 2[\hat{\theta} - E(\hat{\theta})] [E(\hat{\theta}) - \theta] + [E(\hat{\theta}) - \theta]^2\right)
\]

algebra

\[
= E([\hat{\theta} - E(\hat{\theta})]^2) + 2[E(\hat{\theta}) - \theta] E[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]^2
\]

linearity of \( E(\cdot) \)

\[
= V(\hat{\theta}) + 2[E(\hat{\theta}) - \theta] E(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2
\]

\[
= V(\hat{\theta}) + [B(\hat{\theta})]^2
\]
4. Let $Y_1, \ldots, Y_n$ be a random sample of size $n$ from a population with cumulative distribution function

$$F(y) = \begin{cases} 0, & y < 0, \\ y^2/\theta^2, & 0 \leq y \leq \theta, \\ 1, & y > \theta. \end{cases}$$

Let $\hat{\theta} = Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$.

(a) Find the bias of $\hat{\theta}$ as an estimator of $\theta$. \hspace{1cm} (10 pts)

$$F_\theta(y) = [F(y)]^n = \begin{cases} 0, & y < 0, \\ y^{2n}/\theta^{2n}, & 0 \leq y \leq \theta, \\ 1, & y > \theta. \end{cases}$$

$$f_\theta(y) = F_\theta'(y) = \frac{2n}{\theta^{2n}} y^{2n-1}, \quad 0 \leq y \leq \theta$$

$$E(\hat{\theta}) = \int_0^\theta y \frac{2n}{\theta^{2n}} y^{2n-1} \, dy$$

$$= \frac{2n}{\theta^{2n}} \int_0^\theta y^{2n} \, dy = \frac{2n}{\theta^{2n}} \left[ \frac{1}{2n+1} y^{2n+1} \right]_y^{\theta} = \frac{2n}{\theta^{2n}} \frac{\theta^{2n+1}}{2n+1} = \frac{2n}{2n+1} \theta$$

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{2n}{2n+1} \theta - \theta = \left( \frac{2n}{2n+1} - 1 \right) \theta = \frac{-1}{2n+1} \theta$$

(b) Find the variance of $\hat{\theta}$. \hspace{1cm} (10 pts)

$$E(\hat{\theta}^2) = \int_0^\theta y^2 \frac{2n}{\theta^{2n}} y^{2n-1} \, dy$$

$$= \frac{2n}{\theta^{2n}} \int_0^\theta y^{2n+1} \, dy = \frac{2n}{\theta^{2n}} \left[ \frac{1}{2n+2} y^{2n+2} \right]_y^{\theta} = \frac{2n}{\theta^{2n}} \frac{\theta^{2n+2}}{2n+2} = \frac{n}{n+1} \theta^2$$

$$V(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2$$

$$= \frac{n}{n+1} \theta^2 - \left( \frac{2n}{2n+1} \right)^2 = \left( \frac{n}{n+1} - \frac{4n^2}{(2n+1)^2} \right) \theta^2$$

$$= \left( \frac{n(2n+1)^2 - 4n^2(n+1)}{(n+1)(2n+1)^2} \right) \theta^2 = \left( \frac{n}{(n+1)(2n+1)^2} \right) \theta^2$$
5. Suppose that $Y_1, \ldots, Y_n$ be independent and identically distributed with mean $\mu$ and variance $\sigma^2$, where $\mu$ and $\sigma^2$ are unknown. Let $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

(a) Show that
\[
\sum_{i=1}^{n} [(Y_i - \bar{Y})^2] = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.
\]

(b) Show that
\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} [(Y_i - \bar{Y})^2]
\]
is an unbiased estimator of $\sigma^2$.

Note that
\[
E(Y_i) = \mu, \quad V(Y_i) = \sigma^2 \quad \implies \quad E(Y_i^2) = \sigma^2 + \mu^2
\]
\[
E(\bar{Y}) = \mu, \quad V(\bar{Y}) = \frac{\sigma^2}{n} \quad \implies \quad E(\bar{Y}^2) = \frac{\sigma^2}{n} + \mu^2.
\]

Thus by part (a),
\[
E(S^2) = E\left[\frac{1}{n-1} \left( \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \right) \right] = \frac{1}{n-1} \left( \sum_{i=1}^{n} E(Y_i^2) - nE(\bar{Y}^2) \right)
\]
\[
= \frac{1}{n-1} \left[ n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right]
\]
\[
= \frac{1}{n-1} \left[ n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \right] = \frac{1}{n-1} (n-1)\sigma^2 = \sigma^2.
\]
6. Suppose that \(X_1, \ldots, X_m\) is a random sample from a normal distribution with mean \(\mu_x\) and variance \(\sigma^2\), that \(Y_1, \ldots, Y_n\) is a random sample from a normal distribution with mean \(\mu_y\) and variance \(\sigma^2\), and that the two samples are independent (note that the two populations have the same variance). Let \(\bar{X}\) and \(S^2_x\) be the sample mean and variance of the \(X\) sample and let \(\bar{Y}\) and \(S^2_y\) be the sample mean and variance of the \(Y\) sample. Also let

\[
S^2 = \frac{(m-1)S^2_x + (n-1)S^2_y}{m+n-2}.
\]

(a) Name the distribution of \(\frac{(m+n-2)S^2}{\sigma^2}\) and give the value of any parameters. Justify your answer. \(8 \text{ pts}\)

\[
\frac{(m+n-2)S^2}{\sigma^2} = \frac{(m-1)S^2_x + (n-1)S^2_y}{\sigma^2} = \frac{(m-1)S^2_x}{\sigma^2} + \frac{(n-1)S^2_y}{\sigma^2}.
\]

By a theorem from class (since populations are normal)

\[
\frac{(m-1)S^2_x}{\sigma^2} \sim \chi^2_{m-1} \quad \text{and} \quad \frac{(n-1)S^2_y}{\sigma^2} \sim \chi^2_{n-1}.
\]

Because the two samples are independent,

\[
\frac{(m-1)S^2_x}{\sigma^2} \quad \text{and} \quad \frac{(n-1)S^2_y}{\sigma^2} \quad \text{are independent.}
\]

But a sum of independent chi square random variables has a chi square distribution with d.f. equal to the sum of the d.f.'s of the summands, so

\[
\frac{(m+n-2)S^2}{\sigma^2} \sim \chi^2_{m+n-2} \quad \text{(chi square with } m+n-2 \text{ d.f.)}
\]
(b) Find a pivot for $\mu_x$ that is a nontrivial function of $X$, $S$, $\mu_x$, and known constants only, and has a $t$ distribution, and give the degrees of freedom. Justify your answer. Note: Your pivot must not involve $S_x^2$ or $S_y^2$ except through $S^2$. (10 pts)

$$X \sim N(\mu_x, \sigma^2/n) \implies \frac{X - \mu_x}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Also $\bar{X}$ is independent of $S_x^2$ (by theorem) and of $S_y^2$ (since the two samples are independent), so

$$\frac{X - \mu_x}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{and} \quad \frac{(m + n - 2)S^2}{\sigma^2} \sim \chi_{m+n-2}^2 \quad \text{are independent.}$$

But the ratio of a standard normal to the square root of an independent chi square random variable divided by its degrees of freedom has a $t$ distribution. Thus

$$\frac{\bar{X} - \mu_x}{\sigma/\sqrt{n}} \sqrt{\left(\frac{(m + n - 2)S^2}{\sigma^2}\right) / (m + n - 2)} = \frac{\bar{X} - \mu_x}{S/\sqrt{n}} \sim t_{m+n-2} \quad \text{d.f.} = m + n - 2$$
(c) Use your answer to part (b) to construct an exact \((1 - \alpha) \times 100\%\) confidence interval for \(\mu_x\) that depends only on \(\bar{X}, S^2\), and known constants (that might be looked up in a table for example). (6 pts)

\[
1 - \alpha = P\left(-\frac{t_{\alpha/2}}{S/\sqrt{n}} \leq \frac{\bar{X} - \mu_x}{S/\sqrt{n}} \leq \frac{t_{\alpha/2}}{S/\sqrt{n}}\right) = P\left(-\frac{t_{\alpha/2} S/\sqrt{n}}{\bar{X} - \mu_x} \leq \frac{t_{\alpha/2} S/\sqrt{n}}{\bar{X} - \mu_x} \right) = P\left(-\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq -\mu_x \leq -\bar{X} + t_{\alpha/2} S/\sqrt{n}\right) = P\left(\bar{X} + t_{\alpha/2} S/\sqrt{n} \geq -\mu_x \geq \bar{X} - t_{\alpha/2} S/\sqrt{n}\right) = P\left(\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu_x \leq \bar{X} + t_{\alpha/2} S/\sqrt{n}\right).
\]

\[\bar{X} \pm t_{\alpha/2} S/\sqrt{n}\]