Problem Set #1

1. Express $\bar{y}$, $\tilde{y}$, s_y, and IQR_y in terms of $\bar{x}$, $\tilde{x}$, s_x, IQR_x given $y_i = \alpha + \beta x_i$:

The mean and median of y will just be the sample mean and median plugged into the equation:

$$\bar{y} = \alpha + \beta\bar{x}$$
$$\tilde{y} = \alpha + \beta\tilde{x}$$

Plugging in y_i and $\widetilde{y_i}$:

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i - \alpha - \beta\widetilde{x_i})^2}$$
$$s_y = \sqrt{\beta * \frac{1}{n}\sum_{i=1}^{n}(x_i - \widetilde{x_i})^2}$$
$$s_y = \sqrt{\beta}s_x$$

The IQR is only affected by the scale of x but not the translation:

$$IQR_y = \beta(IQR_x)$$

2. Show the interesting optimization interpretations. We can show the first one by differentiating the equation we are trying to minimize and setting it equal to zero.

$$\frac{d}{dx}(\sum_{i=1}^{n}(x_i - \alpha)^2)$$
$$\sum_{i=1}^{n}2(x_i - \alpha) = 0$$

And we have:

$$\sum_{i=1}^{n}\alpha = n\alpha$$

So,

$$\sum_{i=1}^{n}2x_i = 2n\alpha$$
$$\alpha = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Which is the equation for the mean, so $\bar{x}$ is the value α that satisfies the equation:

$$argmin_\alpha \sum_{i=1}^{n}(x_i - \alpha)^2$$

For the median:

$$\frac{d}{dx}(\sum_{i=1}^{n}|x_i - \alpha|)$$

And

$$\frac{d}{dx}|x| = \frac{|x|}{x}$$

Which outputs the sign of the quantity. Therefore, we have

$$\sum_{i=1}^{n} sign(x_i - \alpha) = 0$$

This sum come to zero when half of x_i is to the left of α and half of x_i is to the right of α, which is the definition of the median $\tilde{x}$. Therefore, α is the value that satisfies the equation:

$$argmin_{\alpha} \sum_{i=1}^{n}|x_i - \alpha|$$

3. What can you say about the distribution of the sample if points $\{z_{\frac{k}{n+1}}, x_{(k)}\}$ fall on the line $y = ax + b$ instead of $y = x$?

The line that the points fall on give us insight to the shape and position of the distribution of points. If the points fall on the line $y = ax + b$, we can tell different properties of the distribution of points:

If $a < 1$, we can tell that the distribution has shorter tails than the standard normal distribution. Therefore, the distribution is more clumped towards the middle than the normal distribution.

If $a > 1$, we can tell that the distribution has longer tails than the standard normal distribution. Therefore, the distribution is more spread out than the normal distribution.

If $b < 0$, we can tell that the median of the distribution is shifted in the negative direction.

If $b > 0$, we can tell that the median of the distribution is shifted in the positive direction.

Problem 4 is attached as code.

Problem 5 is attached as code.

6. Compute the following:

a. $P(s_n = N)$
The probability that any given sample value from the sample is equal to N is 1/N since there is a one out of the size of the population chance.

$$P(s_n = N) = \frac{1}{N}$$

This is the same for any value n.

b. $P(\text{the } N^{th} \text{ population unit is in the sample})$
The probability for each sample value to be N is 1/N, so we just add the probabilities:

$$P(\text{the } N^{th} \text{ population unit is in the sample}) = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N}$$

c. $E[s_1]$
The expected value for any sample value would be the mean of the population:

$$E[s_1] = \frac{1}{N} \sum_{i=1}^{N} i$$

d. $P(s_1 = N, s_2 = 1)$
This is equal to one of these events happening and the other one happening without replacement, which we multiply:

$$P(s_1 = N, s_2 = 1) = \frac{1}{N} * \frac{1}{N-1} = \frac{1}{N(N-1)}$$

e. $P(s_i = i, \text{for all } i = 1, \dots, n)$
This is similar to the previous example where we just multiply the probabilities. We do this for the size of the sample:

$$P(s_i = i, \text{for all } i = 1, \dots, n) = \frac{1}{N} * \frac{1}{N-1} * \dots * \frac{1}{N-(n-1)} = \prod_{i=0}^{n-1} \frac{1}{N-i}$$

7a. The sum of the weights should be one:

$$\sum_{i=1}^{n} w_i = 1$$

We can show this:

$$bias\left[\overline{X_n^w}\right] = E\left[\overline{X_n^w}\right] - \mu$$

$$0 = \sum_{i=1}^{n} w_i E[x_i] - \mu$$

$$\mu = \sum_{i=1}^{n} \mu w_i$$
$$1 = \sum_{i=1}^{n} w_i$$

7b. Now, we must minimize the standard error:

$$se = \sqrt{V\left[\overline{X_n^w}\right]}$$

Following equation 13 from lecture 3:

$$V\left[\overline{X_n^w}\right] = V[\sum_{i=1}^{n} w_i x_i] = \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(w_i x_i, w_j x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j cov\left(x_i, x_j\right)$$

We split this summation into two possible cases: i and j are equal or are not equal. We can note that

$$cov(x_i, x_i) = \sigma^2$$
$$cov\left(x_i, x_j\right) = \left(\frac{-\sigma^2}{N-1}\right), i \neq j$$

Therefore, we have

$$V\left[\overline{X_n^w}\right] = \sum_{i=1}^{n} w_i^2 \sigma^2 + (\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left(\frac{-\sigma^2}{N-1}\right), i \neq j)$$

From part a, we know that $\sum_{i=1}^{n} w_i = 1$, so we know $\sum_{i=1}^{n} w_i * \sum_{j=1}^{n} w_j = 1$. Again, we can look at the cases where i and j are equal and not equal:

$$\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j = 1 \; or \; \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j = 1 - \sum_{i=1}^{n} w_i^2$$

Plugging this into our other equation, we get

$$\sum_{i=1}^{n} w_i^2 \sigma^2 + (\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left(\frac{-\sigma^2}{N-1}\right), i \neq j) = \sum_{i=1}^{n} w_i^2 \sigma^2 + (1 - \sum_{i=1}^{n} w_i^2)\left(\frac{-\sigma^2}{N-1}\right)$$

$$= \left(\sum_{i=1}^{n} w_i^2\right)\left(\sigma^2 + \frac{\sigma^2}{N-1}\right) - \frac{\sigma^2}{N-1}$$

With the summation as the only variable.

We will now use the Cauchy-Schwartz inequality, which only works when the variables are linearly dependent. Using this inequality, we get:

$$\left|\sum_{i=1}^{n} w_i\right|^2 \leq \left|\sum_{j=1}^{n} w_j\right|^2 * \sum_{k=1}^{n} 1$$

Where 1 and w_i are linearly dependent. For this to hold, we have

$$< w_1, w_2, \ldots, w_n > = \ a < 1,1,1, \ldots, 1 >$$

For some constant a. This shows that all $w_1, w_2, \ldots, w_n$ are equal, and we know that the sum is equal to one from part a. Therefore,

$$w_i = \frac{1}{n} \ for \ all \ i$$

The most efficient estimate is the sample mean case stated in the problem.