Math 362, Problem set 1

Due 1/31/10

1. (4.1.8) Determine the mean and variance of the mean $\bar{X}$ of a random sample of size 9 from a distribution having pdf $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

   Answer:

   We have that:

   $$\mathbb{E}[\bar{X}] = \mathbb{E}[\frac{1}{n}\sum X_i] = \frac{1}{n} \times n\mathbb{E}[X_i] = \mathbb{E}[X_i]$$

   and

   $$\text{var}(\bar{X}) = \frac{\text{var}(\sum X_i)}{n^2} = \frac{\text{var}(X_i)}{n}.$$

   Using the pdf,

   $$\mathbb{E}[X] = \int_0^1 4x^4 dx = \frac{4}{5}.$$

   and

   $$\mathbb{E}[X^2] = \int_0^1 4x^5 dx = \frac{2}{3},$$

   so

   $$\text{var}(X_i) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75},$$

   and

   $$\text{var}(\bar{X}) = \frac{2}{675}.$$

2. (4.2.25) Let $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$d denote the sample variance of a random sample from a distribution with variance $\sigma^2 > 0$. Since $\mathbb{E}[S^2] = \sigma^2$, why isn't $\mathbb{E}[S] = \sigma$? *Note:* There is a hint in the book that gives it away, but maybe think about it for a second before looking there.

   Answer:

   This follows from the fact that $f(x) = x^2$ is strictly convex (that is $f''(x) = 2 > 0$, and from Jensen's inequality:

   $$\sigma^2 = \mathbb{E}[S^2] > (\mathbb{E}[S]^2),$$

so

$$\mathbb{E}[S] < \sigma.$$

Note: some of you also (cleverly) noticed this follows from the fact that $\text{var}(S) > 0$.

3. (5.1.4) Let $X_1, \ldots, X_n$ be a random sample from the $\Gamma(2, \theta)$ distribution, where θ is unknown. (Recall, look back in chapter 3 if you forget the specifics of the Γ distribution). Let $Y = \sum_{i=1}^{n} X_i$.

   (a) Find the distribution of Y and determine c so that cY is anunbiased estimator of θ.

   (b) If $n = 5$ show that

$$\mathbb{P}\left(9.59 < \frac{2Y}{\theta} < 34.2\right) = 0.95$$

   (c) Using (b), show that if y is the value of Y once the sample is drawn then the interval

$$\left(\frac{2y}{34.2}, \frac{2y}{9.59}\right)$$

   is a 95% confidence interval for Θ.

   (d) Suppose the sample results in the values,

$$44.8079 \quad 1.5215 \quad 12.1929 \quad 12.5734 \quad 43.2305$$

   Based on these data, obtain the point estimate of θ as described in Part (a) and the computed 95% confidence interval in Part (c). What does the confidence interval mean?

Answer:

For (a), $Y \sim \Gamma(2n, \theta)$ by the additive property of the Gamma distribution. Thus $\mathbb{E}[Y] = 2n\theta$ and we should take $c = \frac{1}{2n}$.

For (b), we will simply set up the integral.

$$\begin{aligned}
\mathbb{P}\left(9.59 < \frac{2Y}{\theta} < 34.2\right) &= \mathbb{P}\left(4.795\theta < Y < 17.1\theta\right) \\
&= \int_{4.795\theta}^{17.1\theta} \frac{1}{9!\theta^{10}} x^9 e^{-x/\theta} dx \\
&= \int_{4.795}^{17.1} \frac{u^9}{9!} e^{-u} du \\
&= .9502 \approx .95.
\end{aligned}$$

Note: The important part here is that this *does not depend on* θ.

For (c), note that:

$$\mathbb{P}\left(9.59 < \frac{2Y}{\theta} < 34.2\right) = \mathbb{P}(\frac{2y}{34.2} < \theta < \frac{2y}{9.59}) = .95$$

so the given interval is a 95% confidence interval.

For (d),

$$\theta \approx \frac{y}{2n} = 22.86524,$$

and we have a confidence interval of

$$(6.685, 23.843),$$

meaning that the true value of θ has a 95% chance of landing between these values.

4. (5.1.5) Suppose the number of customers X that enter a store between the hours of 9AM and 10AM follows a Poisson distribution with parameter θ. Suppose a random sample of the number of customers for 10 days results in the values

   $$9 \quad 7 \quad 9 \quad 15 \quad 10 \quad 13 \quad 11 \quad 7 \quad 2 \quad 12$$

   Based on these data obtain an unbiased point estimate of θ. Explain the meaning of this estimate in terms of the number of customers.

   Answer: Since $\mathbb{E}[X] = \theta$ for a Poisson θ random variable, we have a point estimate of 9.5 customers per day. (Of course while it is impossible to actually obtain this on any given day, but it's a fine estimate of the average - it's even perfectly fine if it is the true mean of the distribution.

5. (5.2.2) Obtain the probability that an observationis a potential outlier for the following distributions

   (a) The underlying distribution is normal

   (b) The underlying distribution is *logistic*, in other words it has pdf:

   $$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \qquad -\infty < x < \infty$$

   (c) The underlying distribution is Laplace, the pdf given by:

   $$f(x) = \frac{1}{2}e^{-|x|}, \qquad -\infty < x < \infty.$$

   Answer: For (a), we find (from the back of the book) that

   $$Q3 = 0.675 \qquad Q3 = -0.675.$$

Thus $h = 2.025$, and

$$UF = 2.7 \qquad LF = -2.7.$$

We have $\mathbb{P}(Z > UF) = 0.00346697$, so $\mathbb{P}(LF < Z < UF) = .993066$, so the probability of a potential outlier is ≈ 0.007

For (b), we have that

$$\int_{-\infty}^{Q3} \frac{e^{-x}}{(1+e^{-x})^2} dx = \frac{3}{4},$$

and similar for $Q1$. We thus get $Q3 = \ln(3)$ and $Q1 = -\ln(3)$. Thus $h = 3\ln(3)$, $UF = 4\ln(3)$ and $LF = -4\ln(3)$. We have then, that the

$$\mathbb{P}(LF < X < UF) = \int_{-4\ln(3)}^{4\ln(3)} \frac{e^{-x}}{(1+e^{-x})^2} dx = \frac{40}{41},$$

so the probability of a potential outlier is $\frac{1}{41} \approx .02439$.

For (c) we have that

$$\int_{-\infty}^{Q1} \frac{1}{2} e^{-|x|} = \frac{1}{4}.$$

Solving, we find $Q1 = -\ln(2)$ and likewise $Q3 = \ln(2)$. Thus $h = 3\ln(2)$, and we have $UF = 4\ln(2)$ and $LF = -4\ln(2)$. Since

$$\int_{-4\ln(2)}^{4\ln(2)} \frac{1}{2} e^{-|x|} dx = \frac{15}{16}.$$

Thus the probability of a potential outlier is $\frac{1}{16} = 0.0625$.

6. (5.2.5) Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from the distribution having pdf $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find $\mathbb{P}(3 \leq Y_4)$.

   Answer:

   We have that $F(x) = \int_0^x e^{-t} dx = 1 - e^{-t}$. Using the formula we derived in class:

$$\mathbb{P}(Y_4 \geq 3) = \int_3^\infty 4(1-e^{-t})^3 e^{-t} dt = 1 - (1-e^{-3})^4.$$

7. (5.2.12) Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from a distribution having the pdf $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Show that $Z_1 = Y_1/Y_2$, $Z_2 = Y_2/Y_3$ and $Z_3 = Y_3$ are mutually independent.

   Answer:

We have that $Y_3 = Z_3$, $Y_2 = Z_2 Z_3$ and $Y_1 = Z_1 Z_2 Z_3$. Thus,

$$J = \begin{vmatrix} z_2 z_3 & z_1 z_3 & z_1 z_2 \\ 0 & z_3 & z_2 \\ 0 & 0 & 1 \end{vmatrix} = z_2 z_3^3.$$

We have

$$\begin{aligned} f_{Z_1,Z_2,Z_3}(z_1,z_2,z_3) &= f_{Y_1,Y_2,Y_3}(z_1 z_2 z_3, z_2 z_3, z_3) z_2 z_3^2 \\ &= 3!(2z_1 z_2 z_3)(2z_2 z_3)(2z_3)(z_2 z_3^2) \\ &= (2z_1)(4z_2^3)(6z_3^5), \end{aligned}$$

where $0 < z_i < 1$. This is the product of the marginal pdfs of Z_1, Z_2 and Z_3 which shows that the variables are mutually independent.

8. (5.2.21) Let $X_1, X_2, \ldots, X_n$ be a random sample. A measure of spread is Gini's mean difference

$$G = \sum_{j=2}^{n} \sum_{i=1}^{j-1} |X_i - X_j| / \binom{n}{2}$$

   (a) If $n = 10$, find $a_1, \ldots, a_10$ so that $G = \sum_{i=1}^{10} a_i Y_i$, where $Y_1, Y_2, \ldots, Y_10$ are the order statistics of the sample.

   (b) Show that $\mathbb{E}[G] = 2\sigma/\sqrt{\pi}$ if the sample arises from the normal distribution $N(\mu, \sigma^2)$.

   Hint: This looks worse than it is: Think about the fact that $|X_i - X_j|$ is either $X_i - X_j$ or $X_j - X_i$ depending on which is larger.

Answer:

The Y_i appear in the sum G, as they are some X_j. The key observation is that $|Y_i - Y_j| = Y_i - Y_j$ if $i > j$ and $|Y_i - Y_j| = Y_j - Y_i$, if $i < j$. Thus Y_i appears with a positive sign $i - 1$ times, and with a negative sign $n - i$ times. In all, Y_i has coefficient:

$$a_i = \frac{(i-1) - (n-i)}{\binom{n}{2}}.$$

Since $n = 9$, we have that

$$a_i = \frac{2i - 10}{45}$$

for $i = 1, \ldots, 9$..

For (b), note that $\mathbb{E}[G] = \mathbb{E}[|X_i - X_j|]$. Since $X_i - X_j \sim N(0, 2\sigma^2)$ we have that

$$\begin{aligned} \mathbb{E}[|X_i - X_j|] &= 2\int_0^\infty x \frac{1}{\sqrt{(2\pi)(2\sigma^2)}} \exp(-\frac{1}{2}\frac{x^2}{2\sigma^2}) dx \\ &= 2\int_0^\infty \frac{\sigma}{\sqrt{\pi}} e^{-u} du = \frac{2\sigma}{\sqrt{\pi}}. \end{aligned}$$