Math 3070 § 1. **Third Midterm Exam** Name: Solutions
Treibergs $\sigma\tau$ November 16, 2011

Do FOUR of five problems.

1. *Suppose that we have a random sample with $n = 36$ observations $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$ taken from and exponential distribution with $\lambda = .200$. Let $\overline{X}$ denote the sample mean. What sampling distribution does $\overline{X}$ have? Why? What is the standard error $\sigma_{\overline{X}}$? What is the probability that the sample mean $\overline{X}$ will exceed 6.00?*

The mean and standard deviation of an exponential variable with $\lambda = \frac{1}{5}$ is $\mu_X = \sigma_X = \frac{1}{\lambda} = 5$. Since $n = 36 > 30$, by the rule of thumb we may treat the sample average $\overline{X}$ as being an approximately normal variable from $N(\mu_{\overline{X}}, \sigma_{\overline{X}})$, where $\mu_{\overline{X}} = \mu_X = 5$ and the standard error is $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \boxed{\dfrac{5}{6}}$.

To compute the approximate probability, we standardize

$$\begin{aligned} \mathrm{P}(\overline{X} > 6.000) &\approx \mathrm{P}\left(Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{6.000 - 5.000}{5/6} \right) \\ &= \mathrm{P}(Z > 1.2) = \mathrm{P}(Z < -1.2) = \Phi(-1.2) = \boxed{.1151}. \end{aligned}$$

2. *Let X and Y be random variables whose joint pdf is $f(x, y)$. Find the marginal densities $f_X(x)$ and $f_Y(y)$. Are X and Y independent? Why? Find $\text{Cov}(X, Y)$.*

$$f(x,y) = \begin{cases} x + y, & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1; \\ \\ 0, & \text{otherwise.} \end{cases}$$

The marginal density is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \begin{cases} \int_0^1 x + y\,dy = x + \frac{1}{2}, & \text{if } 0 \le x \le 1; \\ \\ 0, & \text{otherwise.} \end{cases}$$

By symmetry, $f_Y(y) = f_X(y)$. X and Y are not independent, because *e.g.*, for $0 \le x, y \le 1$,

$$f_X(x) f_Y(y) = \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right) = xy + \frac{x}{2} + \frac{y}{2} + \frac{1}{4} \ne x + y = f(x,y).$$

The expectation

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x\left(x + \frac{1}{2}\right)\,dx = \int_0^1 \left(x^2 + \frac{x}{2}\right)\,dx = \left[\frac{x^3}{3} + \frac{x^2}{4}\right]_0^1 = \frac{7}{12}.$$

By symmetry, $\mathrm{E}(Y) = \mathrm{E}(X)$. Expected XY is

$$\mathrm{E}(XY) = \int_0^1 \int_0^1 xy(x+y)\,dy\,dx = \int_0^1 \frac{x^2}{2} + \frac{x}{3}\,dx = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

The covariance is thus $\text{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \dfrac{1}{3} - \dfrac{7}{12} \cdot \dfrac{7}{12} = \boxed{-\dfrac{1}{144}}$.

3. *The article "An Evaluation of Football Helmets Under Impact Conditions" (Amer. J. Sports Medicine, 1984) reports that when each football helmet in a random sample of 27 suspension-type helmets was subjected to a certain impact test, 18 showed damage. Let p denote the proportion of of all helmets of this type that would show damage when tested in the prescribed manner. Calculate a 99% two-sided confidence interval for p. What sample size would be required for the width of a 99% CI to be at most .10, irrespective of $\hat{p}$?*

The estimator is $\hat{p} = \frac{18}{27} = \frac{2}{3}$. Since $n\hat{p} = 18$ and $n\hat{q} = 9$ we use the score confidence interval that is valid even for small sample sizes. For a $99\% = 1 - \alpha$ two sided bound, we need the critical value $z_{\alpha/2} = z_{.005} = 2.576$ from Table A5. The CI is

$$\frac{\hat{p} + \dfrac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z_{\alpha/2}^2}{4n^2}}}{1 + \dfrac{z_{\alpha/2}^2}{n}} = \frac{\dfrac{2}{3} + \dfrac{(2.576)^2}{2 \cdot 27} \pm 2.576\sqrt{\dfrac{\frac{2}{3} \cdot \frac{1}{3}}{27} + \dfrac{(2.576)^2}{4 \cdot 27^2}}}{1 + \dfrac{(2.576)^2}{27}} = \boxed{(.422, .846)}$$

We use the width of the traditional interval to estimate n since we expect it to be large. Since $4\hat{p}\hat{q} \le 1$ for all $\hat{p}$,

$$w = 2z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \le \frac{z_{\alpha/2}}{\sqrt{n}}$$

which is less than .1 if

$$n \ge \frac{z_{\alpha/2}^2}{(.1)^2} = \frac{(2.576)^2}{(.1)^2} = 663.5776.$$

The 99% confidence interval will have width at most .10 for $n = \boxed{664}$.

[The study actually reported 37 damaged helmets out of 45 tested.]

4. *Let $0 < p < 1$. A Bernoulli random variable X takes the values $X \in \{0, 1\}$ and has the pmf $p(0) = 1 - p$, $p(1) = p$ and $p(x) = 0$ otherwise. Take a random sample X, Y of two* Bernoulli(p) *variables. Consider the family of statistics defined for $0 < \alpha < 1$ by*

$$\hat{\theta}_\alpha = \alpha X + (1 - \alpha)Y.$$

Show that the statistics $\hat{\theta}_\alpha$ are unbiased estimators for p. Determine the standard errors $s_{\hat{\theta}_\alpha}$ of the statistics $\hat{\theta}_\alpha$. Among the $\hat{\theta}_\alpha$'s with $0 < \alpha < 1$, which is the best estimator for p and why?

If $X \sim$ Bernoulli(p) then $\mathrm{E}(X) = p$ and $\mathrm{V}(X) = pq$. Using linearity of expectation, the statistic is an unbiased estimator for p because

$$\mathrm{E}(\hat{\theta}_\alpha) = \mathrm{E}(\alpha X + (1 - \alpha)Y) = \alpha\mathrm{E}(X) + (1 - \alpha)\mathrm{E}(Y) = \alpha p + (1 - \alpha)p = p.$$

Because of independence of X and Y, the variance is

$$\mathrm{V}(\hat{\theta}_\alpha) = \mathrm{V}(\alpha X + (1 - \alpha)Y) = \alpha^2\mathrm{V}(X) + (1 - \alpha)^2\mathrm{v}(Y) = \left[\alpha^2 + (1 - \alpha)^2\right] pq.$$

Thus the standard error is

$$\sigma_{\hat{\theta}_\alpha} = \sqrt{\alpha^2 pq + (1 - \alpha)^2 pq}.$$

The best estimator among these unbiased estimators is the one with least variance. As the variance is a positive quadratic function in α, its minimum is where the derivative vanishes,

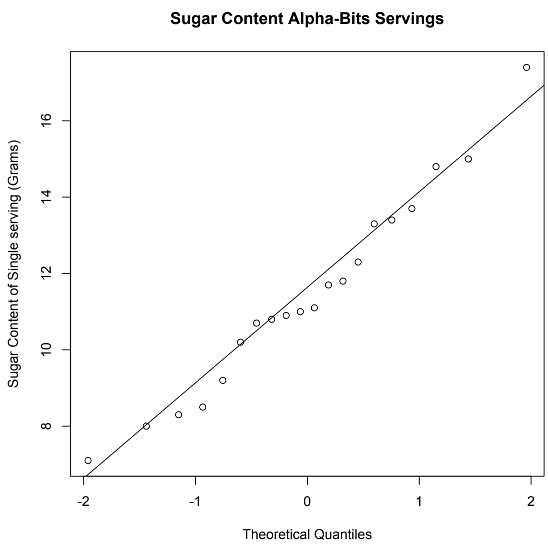
$$\frac{d}{d\alpha}\mathrm{V}(\hat{\theta}_\alpha) = \left[2\alpha - 2(1 - \alpha)\right] pq = 0$$

or at $\alpha = \dfrac{1}{2}$. The best estimator is thus $\boxed{\hat{\theta}_{1/2}}$.

5. *A National Institute of Health study measured the sugar content (in grams) of a random sample of 20 similar single servings of Alpha-Bits cereal. The data is entered into* **R**©*:*

```
> X <- scan()
1:   7.1   8.0   8.3   8.5   9.2  10.2  10.7  10.8  10.9  11.0
11: 11.1  11.7  11.8  12.3  13.3  13.4  13.7  14.8  15.0  17.4
21:
Read 20 items
> mean(X); sd(X)
[1] 11.46
[1] 2.616828
```

Find a 95% lower confidence bound for the mean sugar content of a single serving. Under what assumptions is your confidence bound valid? Based on the accompanying **R**© *generated normal PP-Plot, comment on the validity of your assumptions.*



Since $n \le 40$, there are a small number of observations so we use the t-distribution based CI. The degrees of freedom is $\nu = n - 1 = 20 - 1 = 19$. The one-sided critical value for confidence level $.95 = 1 - \alpha$ is $t_{\alpha,\nu} = t_{.05,19} = 1.729$ from Table A5. The lower confidence bound on μ, the population mean, is using values from the printout,

$$\bar{X} - t_{\alpha,\nu} \frac{S}{\sqrt{n}} = 11.46 - 1.729 \cdot \frac{2.616828}{\sqrt{20}} = \boxed{10.45}.$$

Thus with 95% confidence, $10.45 < \mu$.

This confidence bound is valid provided that the sample is taken from an approximately normal distribution. The normal PP-plot indicates that the observations line up nicely with the theoretical quantiles, indicating that the data is plausibly normal. (In fact, a normal random number generator was used to generate these data based on the reported $\overline{X}$ and S from the study.)