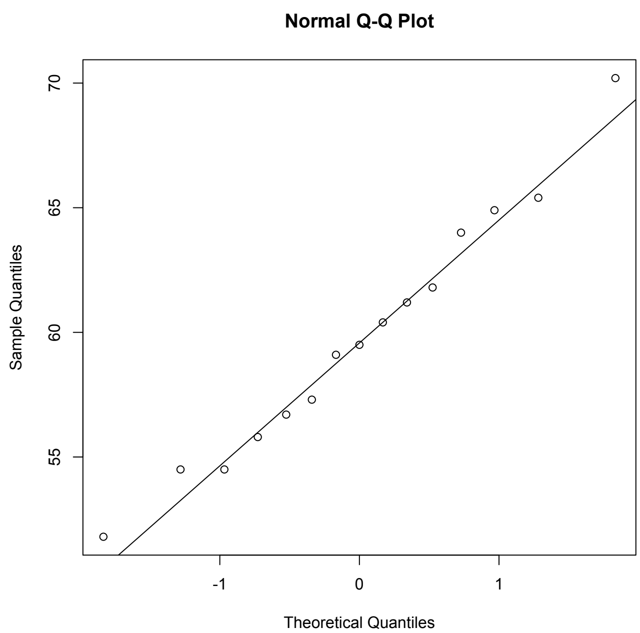
Math 3070 § 1. **Third Midterm Exam** Name: Solutions
Treibergs σπ November 15, 2012

1. *In a study of water absorbtion properties of cotton fabric, the following measurements were made of percentage of water pickup rates for 15 samples. Find a 95% upper confidence bound for μ the mean absorbtion percentage. What assumptions are needed on the data in order that your confidence interval be valid? Comment on how well the data satisfies the assumptions.*

```
> x <- scan()
1: 51.8 59.5 59.1 61.8 61.2 55.8 57.3 64.9 65.4 54.5 54.5 60.4 64.0 70.2 56.7
16:
Read 15 items
 > mean(x); sd(x)
[1] 59.80667
[1] 4.943317
> qqnorm(x); qqline(x)
```



As this is a small sample with σ unknown, we use the confidence interval for the mean coming from the t-distribution. We have $\nu = n-1 = 14$ degrees of freedom and a confidence level of $\alpha = .05$, so the critical value for a one-sided interval is $t_{\alpha,\nu} = t_{.05,14} = 1.761$ from Table A5. Thus the one-sided upper CI for μ is

$$\mu < \overline{X} + t_{\alpha,\nu}\frac{s}{\sqrt{n}} = 59.807 - 1.761 \cdot \frac{4.933}{\sqrt{15}} = \boxed{62.05}.$$

For the t-distribution based interval to be valid, we need to assume that the underlying distribution is approximately normal. Judging from the normal $P-P$ plot, the points line up nicely, so there is no indication that the data is not normal.

2. *A 2002 study in Journal of Environmental Engineering looked at landfills containing demolition waste. It found that 33 of 42 such sites contained detectable levels of chromium. Find a 90% lower confidence interval on the probability that such a site contains detectable levels of chromium. How big should the sample size n be so that the lower bound is within .05 of $\hat{p}$?*

The estimator for p is $\hat{p} = \frac{X}{n} = \frac{33}{42} = \frac{11}{14} = 0.2619048$ where $n = 42$. $\hat{q} = \frac{9}{42}$ so that $n\hat{q} = 9$. Therefore the condition $n\hat{p} \ge 10$ and $n\hat{q} \ge 10$ fails and we must use the unrestricted score interval. We want the confidence level $\alpha = .10$ so for a one-sided interval the critical value is $z_\alpha = z_{0.10} = 1.282$ from Table A5. The one-sided lower confidence interval is

$$p > \frac{\hat{p} + \frac{z_\alpha^2}{2n} - z_\alpha\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_\alpha^2}{4n^2}}}{1 + \frac{z_\alpha^2}{n}} = \frac{\frac{11}{14} + \frac{(1.282)^2}{2 \cdot 42} - (1.282)\sqrt{\frac{\frac{11}{14}\frac{3}{14}}{42} + \frac{(1.282)^2}{4 \cdot 42^2}}}{1 + \frac{(1.282)^2}{42}} = \boxed{0.695}.$$

One could use Formula 7.12 with $w = .10 = 2 \cdot (.05)$, or its simplified approximation, but we expect that for small width, the sample size will be large and the traditional interval would be appropriate. Thus, we want the lower bound L within .05 of $\hat{p}$ or $|L - \hat{p}| \le .05$. The lower bound would be

$$p > L = \hat{p} - z_\alpha\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

so

$$.05 \ge |L - \hat{p}| = z_\alpha\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

or

$$n \ge \frac{z_\alpha^2 \hat{p}\hat{q}}{(.05)^2}.$$

This is implied if we take

$$n = \frac{.25 z_\alpha^2}{(.05)^2} = 164.4.$$

because $.25 \ge \hat{p}\hat{q}$ for any $\hat{p} = 1 - \hat{q}$. Thus a sample size of $\boxed{n = 165}$ would give the desired width.

3. *Suppose that an urn contains three red balls, one white ball and two blue balls. If three balls are drawn at random without replacement, let X be the number of red balls and Y the number of white balls drawn. Then the joint probability mass function is given by $p(x, y)$. Let $Z = \max\{X, Y\}$ be the maximum of the two numbers. Find the pmf for the random variable Z. What is $E(Z)$? Are X and Y independent? Why? Find Cov(X, Y).*

$p(x,y)$		x = 0	1	2	3	$p_Y(y)$
y	0	0	.15	.30	.05	.5
	1	.05	.30	.15	0	.5
	$p_X(x)$	.05	.45	.45	.05	

The pmf for $Z = \max(X, Y)$ is given for all possible values of $z \in \{0, 1, 2, 3\}$. $p_Z(0) = P(\{Z = 0\}) = P(\{(0, 0)\}) = p(0, 0) = 0$. $p_Z(1) = P(\{Z = 1\}) = P(\{(1, 0), (0, 1), (1, 1)\}) = p(1, 0) + p(0, 1) + p(1, 1) = .5$. $p_Z(2) = P(\{Z = 2\}) = P(\{(2, 0), (2, 1)\}) = p(2, 0) + p(2, 1) = .45$. $p_Z(3) = P(\{Z = 3\}) = P(\{(3, 0), (3, 1)\}) = p(3, 0) + p(3, 1) = .05$. So

z	0	1	2	3
$p_Z(z)$	0	.50	.45	.05

Thus

$$E(Z) = \sum_{z=0}^{3} z p_Z(z) = 0 \cdot 0 + 1 \cdot (.50) + 2 \cdot (.45) + 3 \cdot (.05) = \boxed{1.55}.$$

Alternatively use the formula $E(Z) = \sum_{x=0}^{3} \sum_{y=0}^{1} \max(x, y) p(x, y)$.

The marginal probabilities are the row and column sums. X and Y **are not independent** because, *e.g.*, $0 = p(0, 0) \neq p_X(0)p_Y(0) = (.05)(.5) = .025$.

We have

$$E(X) = \sum_{x=0}^{3} x\, p_X(x) = 0 \cdot (.05) + 1 \cdot (.45) + 2 \cdot (.45) + 3 \cdot (.05) = 1.5,$$

$$E(Y) = \sum_{y=0}^{1} y\, p_Y(y) = 0 \cdot (.5) + 1 \cdot (.5) = .5,$$

$$E(XY) = \sum_{x=0}^{3} \sum_{y=0}^{1} xy\, p(x, y) = 1 \cdot 1 \cdot (.30) + 2 \cdot 1 \cdot (.15) = .60.$$

Thus, using the short cut formula

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = .60 - (1.5)(.5) = \boxed{-.15}.$$

4. *Suppose that the time in seconds a snowflake survives after landing on a sleeve is a random variable X with the probability density function with parameter $\theta > 0$ given by $f(x;\theta)$. Compute the expected value $E(X)$. Suppose that X, Y is a random sample of snowflake survival times taken from this distribution. Let α be a real number. Show that $\hat{\theta}_\alpha = \alpha X + (3-\alpha)Y$ is an unbiased estimator for θ. Compute the standard error of $\hat{\theta}_\alpha$. Among the $\hat{\theta}_\alpha$, which one gives the best estimator for θ? Why?*

$$f(x;\theta) = \begin{cases} \dfrac{2}{\theta} - \dfrac{2}{\theta^2}x, & \text{if } 0 \le x \le \theta; \\ \\ 0, & \text{otherwise.} \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^\theta x \left(\frac{2}{\theta} - \frac{2}{\theta^2}x \right) dx = \left[\frac{1}{\theta}x^2 - \frac{2}{3\theta^2}x^3 \right]_0^\theta = \boxed{\frac{\theta}{3}}.$$

Both X and Y are independent with the same distribution. Thus, the estimators $\hat{\theta}_\alpha$ are unbiased because

$$E(\hat{\theta}_\alpha) = E(\alpha X + (3-\alpha)Y) = \alpha E(X) + (3-\alpha)E(Y) = \frac{\alpha\theta}{3} + \frac{(3-\alpha)\theta}{3} = \theta.$$

We have

$$E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)\, dx = \int_0^\theta x^2 \left(\frac{2}{\theta} - \frac{2}{\theta^2}x \right) dx = \left[\frac{2}{3\theta}x^3 - \frac{2}{4\theta^2}x^4 \right]_0^\theta = \boxed{\frac{\theta^2}{6}}$$

so

$$V(X) = E(X^2) - E(X)^2 = \frac{\theta^2}{6} - \frac{\theta^2}{9} = \frac{\theta^2}{18}.$$

Thus, since X and Y are independent,

$$V(\hat{\theta}_\alpha) = V(\alpha X + (3-\alpha)Y) = \alpha^2 V(X) + (3-\alpha)^2 V(Y) = \left[\alpha^2 + (3-\alpha)^2\right] \frac{\theta^2}{18}.$$

Hence the standard error is

$$\sigma_{\theta_\alpha} = \left[\alpha^2 + (3-\alpha)^2\right]^{\frac{1}{2}} \frac{\theta}{3\sqrt{2}}.$$

which is minimum when $\alpha = 1.5$. Thus the best of these estimators is the one with the smallest standard error, namely $\boxed{\hat{\theta}_{1.5}}$.

5. *Electronic books manufactured by Cedar City Readers have a mean lifetime of 6.75 years and a standard deviation of 0.90 years, while the Maeser Laser has a mean lifetime of 6.40 years and a standard deviation of 1.40 years. Let* $\overline{X}$ = *the average lifetime for a random sample of 36 Cedar Readers and* $\overline{Y}$ = *the average lifetime for a sample 49 Maeser Lasers. How is the random variable* $\overline{X} - \overline{Y}$ *distributed? What are its mean* $\mu_{\overline{X}-\overline{Y}}$ *and standard deviation* $\sigma_{\overline{X}-\overline{Y}}$*? Why? What is the probability that the average lifetime of a sample of 36 Cedar Readers will be at least one year more than the average lifetime of a sample of 49 Maeser Lasers?*

Because we assume that $X_1, \ldots, X_{36}$ are IID random variables from a distribution with $\mu_X = 6.75$ and $\sigma_X = 0.90$, and that $n_X = 36 > 30$, then the mean is approximately normally distributed by the rule of thumb for the Central Limit Theorem, $\overline{X} \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n_X}})$.

Because $Y_1, \ldots, Y_{49}$ are IID random variables from a distribution with $\mu_Y = 6.40$ and $\sigma_Y = 1.40$, and that $n_Y = 49 > 30$, then the mean is also approximately normally distributed $\overline{Y} \sim N(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}})$. Finally because the X_i's are independent of the Y_j's so $\overline{X}$ is independent of $\overline{Y}$, we use the formula for linear combinations to find that

$$\mu_{\overline{X}-\overline{Y}} = \mu_X - \mu_Y,$$
$$\sigma^2_{\overline{X}-\overline{Y}} = V\left(\overline{X} - \overline{Y}\right) = (1)^2 V(\overline{X}) + (-1)^2 V(\overline{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}.$$

As the difference of two independent normal variables is also normal, we have

$$\overline{X} - \overline{Y} \sim N(\mu_{\overline{X}-\overline{Y}}, \sigma_{\overline{X}-\overline{Y}}).$$

Substituting values

$$\mu_{\overline{X}-\overline{Y}} = \mu_X - \mu_Y = 6.75 - 6.40 = \boxed{.35},$$
$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = \sqrt{\frac{(0.90)^2}{36} + \frac{(1.40)^2}{49}} = \boxed{0.25}.$$

Finally, standardizing

$$\begin{aligned} P(\overline{X} \ge \overline{Y} + 1) &= P(\overline{X} - \overline{Y} \ge 1) \\ &= P\left(Z = \frac{\overline{X} - \overline{Y} - \mu_{\overline{X}-\overline{Y}}}{\sigma_{\overline{X}-\overline{Y}}} \ge \frac{1 - .35}{.25} = 2.60\right) \\ &= 1 - P(Z < 2.60) = 1 - \Phi(2.60) = \Phi(-2.60) = \boxed{.0047}. \end{aligned}$$