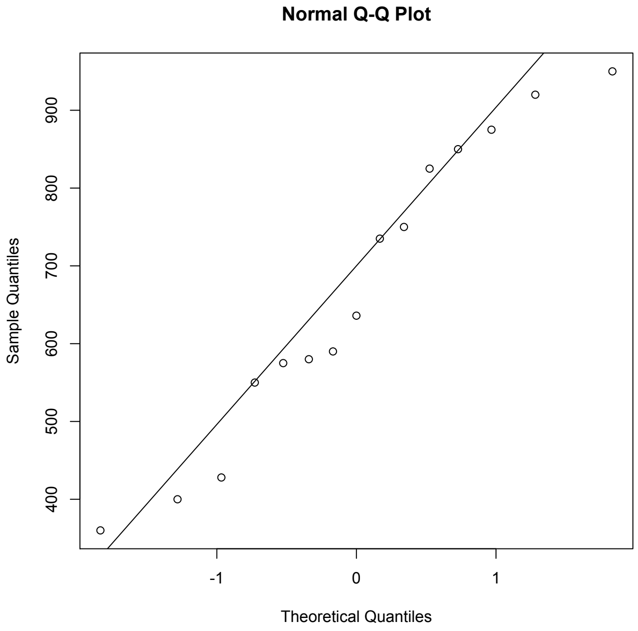
Math 3070 § 1. **Third Midterm Exam** Name:________________
Treibergs November 20, 2013

1. *In a 2001 study in the Journal of Structural Engineering, the nominal shear strength in kN was measured for 15 prestressed concrete beams. Find a 99% upper confidence interval for μ the mean shear strength. What assumptions are needed on the data in order that your confidence interval be valid? Comment on how well the data satisfies the assumptions.*

```
> ShearStrength <- scan()
1: 580 400 428 825 850 875 920 550 575 750 636 360 590 735 950
16:
Read 15 items
> c(mean(ShearStrength),  sd(ShearStrength))
[1] 668.2667   192.0891
> qqnorm(ShearStrength); qqline(ShearStrength)
```



We have a small sample with $n \le 40$ and don't know σ. If the data is approximately normal, then the appropriate interval is based on the t-distribution with $\nu = n - 1 = 14$ degrees of freedom. For a significance level of $\alpha = .01$, the critical value for the one-sided interval is is $t_{\alpha,\nu} = t_{.01,14} = 2.624$ from Table A5. Then reading from the output, the upper CI on μ is given by

$$\mu < \bar{X} + t_{\alpha,\nu}\frac{S}{\sqrt{n}} = 668.2667 + 2.624 \cdot \frac{192.0891}{\sqrt{15}} = \boxed{798.4.}$$

The t-distribution based interval is appropriate if the data is approximately normally distributed. Looking at the QQ-plot, in view of the small n, the points are lining up very nicely. There is no strong indication that normality is not plausible. The points have a slight "S shape" suggesting that the distribution has light tails, but with such small n this is as linear as a QQ-plot gets.

2. *Let X_1, X_2, and X_3 be a random sample taken from a Bernoulli distribution (X_i equals one with probability p and equals zero with probability $1-p$). Consider the statistic*

$$\widehat{p} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3.$$

What is the sampling distribution of $\widehat{p}$? Why? What is its standard error $\sigma_{\widehat{p}}$? Why? Suppose that we consider a second statistic

$$\widehat{\vartheta} = 0.3X_1 + 0.3X_2 + 0.4X_3.$$

Show that $\widehat{p}$ and $\widehat{\vartheta}$ are both unbiased estimators of p. Which of $\widehat{p}$ and $\widehat{\vartheta}$ is the better estimator for p? Why?

The sum of three Bernoulli variables $T = X_1 + X_2 + X_3$ is a binomial variable where T is the number of successes out of $n = 3$ independent trials, and the probability of a success is p. Since T takes values in $\{0, 1, 2, 3\}$, the statistic $\widehat{p}$ takes values in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ and the corresponding sampling distribution is given by the PMF

x	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$p(x)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

Because the binomial variable has the variance $V(T) = 3p(1-p)$, the standard error is

$$\sigma^2_{\widehat{p}} = V\left(\frac{T}{3}\right) = \frac{3p(1-p)}{3^2} \qquad \text{so} \qquad \sigma_{\widehat{p}} = \boxed{\sqrt{\frac{p(1-p)}{3}}}.$$

Both statistics have the form

$$c_1X_1 + c_2X_2 + c_3X_3 \qquad \text{where} \qquad c_1 + c_2 + c_3 = 1.$$

Thus the expectation

$$E(c_1X_1 + c_2X_2 + c_3X_3) = c_1E(X_1) + c_2E(X_2) + c_3E(X_3) = c_1p + c_2p + c_3p = p.$$

Hence both such statistics are unbiased.

The best statistic among two unbiased estimators for p is the one with the least variance. Because X_i are independent and $V(X_i) = p(1-p)$,

$$\begin{aligned} V(\widehat{\vartheta}) &= V(0.3X_1 + 0.3X_2 + 0.4X_3.) \\ &= (0.3)^2V(X_1) + (0.3)^2V(X_2) + (0.4)^2V(X_3) \\ &= (.09 + .09 + .16)p(1-p) = .34p(1-p). \end{aligned}$$

Finally, since $.34p(1-p) = V(\widehat{\vartheta}) > V(\widehat{p}) = \frac{1}{3}p(1-p)$, we see that $\widehat{p}$ is the better estimator.

3. *A 2001 article in* Environmental and Resource Economics *reported on the results of a survey in which Scottish voters were asked if they would be willing to pay additional taxes in order to restore the Affric forest. Out of 175 who responded, 56 said that they would be willing to pay. Using the procedure recommended in the text, find a 90% two sided confidence interval on the proportion willing to pay. How big should the sample size n be to specify the proportion within $\pm .03$ of $\hat{p}$?*

   The score interval is the recommended CI on proportion. The estimate for p, the proportion of Scottish voters in favor of the tax increase is $\widehat{p} = \frac{X}{n} = \frac{56}{175} = .32$, where $n = 175$ is the number polled, and $X = 56$ the number of those in favor. Thus $\widehat{q} = 1 - \widehat{p} = .68$. For the $\alpha = .10$ confidence level, the two-sided critical value needed in the computation is $z_{\alpha/2} = z_{.05} = 1.645$ from Table A5. The interval is given by

$$\frac{\widehat{p} + \dfrac{z_{\frac{\alpha}{2}}^2}{2n} \mp z_{\frac{\alpha}{2}} \sqrt{\dfrac{\widehat{pq}}{n} + \dfrac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \dfrac{z_{\frac{\alpha}{2}}^2}{n}} = \frac{.32 + \dfrac{(1.645)^2}{2(175)} \mp 1.645 \sqrt{\dfrac{(.32)(.68)}{175} + \dfrac{(1.645)^2}{4(175)^2}}}{1 + \dfrac{(1.645)^2}{175}}$$

   This interval works out to be $\boxed{(0.265, 0.380)}$.

   To get an interval of width $w = .06$ any of the three estimates for n are acceptable. We may use the number calculation which assumes the same values of $\widehat{p}$ as the sample here.

$$\begin{aligned} n &= \frac{2z_{\frac{\alpha}{2}}^2 \widehat{pq} - z_{\frac{\alpha}{2}}^2 w^2 + \sqrt{4z_{\frac{\alpha}{2}}^2 \widehat{pq}(\widehat{pq} - w^2) + w^2 z_{\frac{\alpha}{2}}^4}}{w^2} \\ &= \tfrac{2(1.645)^2(.32)(.68) - (1.645)^2(.06)^2 + \sqrt{4(1.645)^2(.32)(.68)((.32)(.68) - (.06)^2) + (.06)^2(1.645)^4}}{(.06)^2} \end{aligned}$$

   which equals 651.9. Thus we take at least $\boxed{n = 652}$. The simpler formula works reasonably well

$$n \approx \frac{4z_{\frac{\alpha}{2}}^2 \widehat{pq}}{w^2} = \frac{4(1.645)^2(.32)(.68)}{(.06)^2} = 654.3$$

   which says use $n = 655$. The simplest, most conservative formula uses the fact that $1 \geq 4\widehat{pq}$ which estimates n without knowing $\widehat{p}$ gives

$$n \approx \frac{z_{\frac{\alpha}{2}}^2}{w^2} = \frac{(1.645)^2}{(.06)^2} = 751.7$$

   which says to use $n = 752$.

4. *Two different roads merge to form the Bangerter Highway. Suppose that during the noon hour, the number of cars coming from each road are random variables X and Y with population means $\mu_X = 800$ and $\mu_Y = 1000$, standard deviations $\sigma_X = 16$ and $\sigma_Y = 25$ and covariance $\mathrm{Cov}(X,Y) = 80$. What is the expected total number of cars entering the Bangerter Highway? Assuming that X and Y are normally distributed, what is the probability that the total number exceeds 1850 cars?*

   The total number of cars entering Bangerter Highway during a random noonhour is $T = X + Y$. Its expectation is

$$E(T) = E(X) + E(Y) = 800 + 1000 = \boxed{1800}.$$

   To compute the variance, we use the formula for non-independent variables

$$\begin{aligned}\sigma_T^2 &= V(1\cdot X + 1\cdot Y)\\ &= 1\cdot 1\cdot V(X) + 2\cdot 1\cdot 1\cdot \mathrm{Cov}(X,Y) + 1\cdot 1\cdot V(Y)\\ &= (16)^2 + 2\cdot 80 + (25)^2 = 1041.\end{aligned}$$

   If X and Y are normal, then so is $X+Y$. Thus, standardizing

$$\begin{aligned}P(T > 1850) &= P\left(z = \frac{T-\mu_T}{\sigma_T} > \frac{1850-1800}{\sqrt{1041}} = 1.5497\right)\\ &= P(z < -1.5497) = \Phi(-1.5497) = \boxed{.0606}\end{aligned}$$

   Since $-1.5497 = (.03)(-1.54) + (.97)(-1.55)$ we interpolate $\Phi(-1.5497) \approx (.03)\Phi(-1.54) + (.97)\Phi(-1.55) = (.03)(.0618) + (.97)(.0606) = .0606$.

5. *Suppose that the the continuous random variables X and Y have the joint density function given by*

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

   Find the marginal density functions $f_X(x)$, $f_Y(y)$. Are X and Y independent? Why? Find the covariance $\mathrm{Cov}(X,Y)$.

   The marginal density is given by

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

   By symmetry, $f_Y(y) = f_X(y) = y + \frac{1}{2}$ for $0 \le y \le 1$ and $f_Y(y) = 0$ otherwise.

   X and Y are not independent since $f(x,y)$ does not equal $f_X(x)f_Y(y)$. For example

$$.3 = .1 + .2 = f(.1,.2) \ne f_X(.1)f_Y(.2) = (.1+.5)(.2+.5) = .42.$$

   The covariance is given by the shortcut formula $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Computing

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

By symmetry, $E(Y) = E(X) = \frac{7}{12}$. Finally

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x,y)\, dx\, dy \\ &= \int_{y=0}^{1} \int_{x=0}^{1} xy(x+y)\, dx\, dy \\ &= \int_{y=0}^{1} \left[\frac{x^3 y}{3} + \frac{x^2 y^2}{2} \right]_{x=0}^{1} dy \\ &= \int_{y=0}^{1} \left(\frac{y}{3} + \frac{y^2}{2} \right) dy \\ &= \left[\frac{y^2}{6} + \frac{y^3}{6} \right]_{y=0}^{1} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

Finally

$$\operatorname{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \left(\frac{7}{12} \right)^2 = \boxed{-\frac{1}{144}}.$$