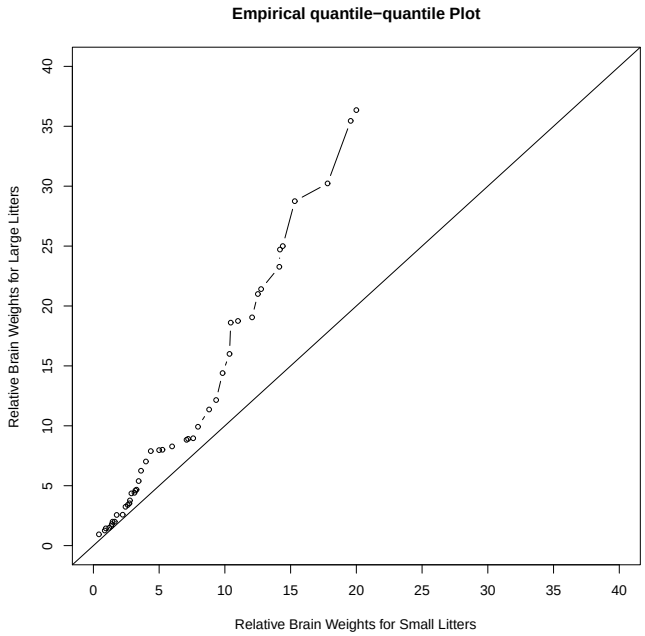
Stat 641

Solutions for Assignment 5

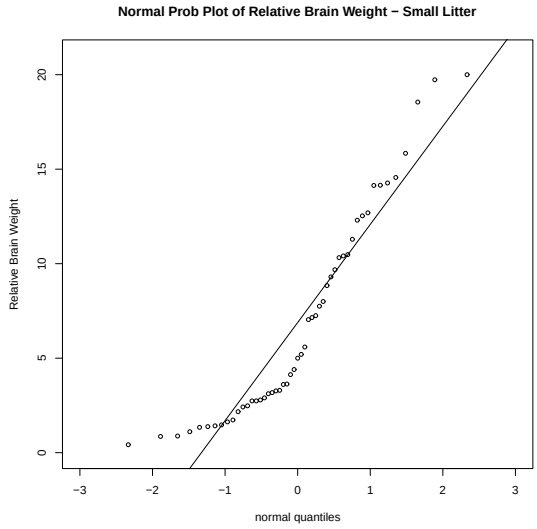
I. (6 points) The process distribution appears to be symmetric with both tails much longer (heavier) than a normal distribution.

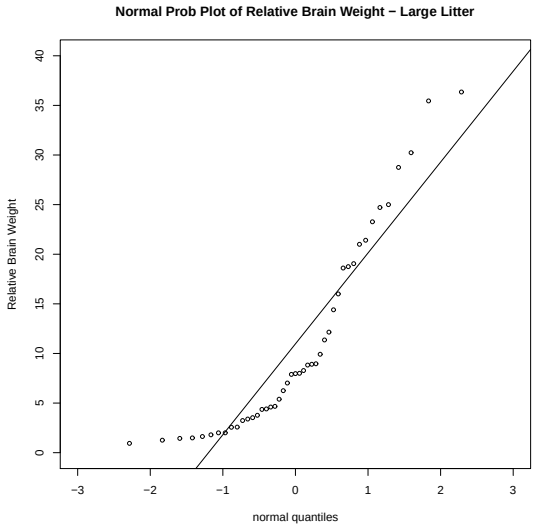
II. (15 points)

1.



2.





3. Based on the graphs, the relative brain weights for both the large and small litters appear to be non-normally distributed with right skewed distributions. From the qqplot, it appears that the relative brain weights for the large litters are much larger than the weights from the small litters based on having all the plotted points being above the 45° line. Furthermore, the plotted points in the qqplot are relatively close to a line which would indicate the the distributions for the two litter sizes may be of the same family but with different parameter values. This is reenforced by the very similar normal reference plots for the two litter sizes.

III. (15 points) The probability can be directly calculated using the cdf of Y, F_Y:

$$p_o = P[\text{outlier}] = F_Y\left[Q(.25) - 1.5(IQR)\right] + 1 - F_Y\left[Q(.75) + 1.5(IQR)\right]$$

- If F_Y is from a location/scale family and Z is the standard member of the family then:

$$p_o = P[\text{outlier}] = F_Z\left[Q_Z(.25) - 1.5(IQR_Z)\right] + 1 - F_Z\left[Q_Z(.75) + 1.5(IQR_Z)\right]$$

- For a distribution which is symmetric about 0,

$$Q(0.75) = -Q(0.25) \quad \Rightarrow \quad IQR = Q(0.75) - Q(0.25) = -2Q(0.25).$$

Thus, for a distribution which is symmetric about 0, we can simplify the probability that a randomly selected observation is an outlier as follows:

$$\begin{aligned} p_o &= P[Y < Q(0.25) - 1.5IQR] + P[Y > Q(0.75) + 1.5IQR] \\ &= 2P[Y < Q(0.25) - 1.5IQR] \quad \text{when the given distribution is symmetric about 0.} \\ &= 2P[Y < 4Q(0.25)]. \end{aligned}$$

1. **exponential** has cdf $F(t) = 1 - e^{-t/\beta}$ and quantile function, $Q(u) = -\beta log(1-u)$, hence β is a scale parameter. Therefore, we have

   $Q(.25) = -\beta log(.75),\ Q(.75) = -\beta log(.25),\ \Rightarrow\ IQR = \beta log(3) \Rightarrow$

   $p_o = P[Y < -\beta log(.75) - 1.5\beta log(3)] + P[Y > -\beta log(.25) + 1.5\beta log(3)]\ \Rightarrow$

   $p_o = P[Y/\beta < -1.36] + 1 - P[Y/\beta < 3.0342] = 0 - e^{-3.0342} = .048$

2. **Weibull** - See discussion in HO 8 - The probability of an outlier depends on the value of γ.

3. **uniform on (0,1)** has cdf $F(y) = y$ for $0 < y < 1\ \Rightarrow\ Q(u) = u\ \Rightarrow\ Q(.25) = .25,\ Q(.75) = .75,\ IQR = .5$

   $p_o = P[Y < .25 - (1.5)(.5)] + P[Y > .75 + (1.5)(.5)] = P[Y < -.5] + 1 - P[Y \leq 1.5] = 0 + 1 - 1 = 0$

4. **Normal:** Let Z have a $N(0,1)$ distribution. Then the probability of an outlier for Y having a $N(\mu, \sigma^2)$ distribution is the same as the probability of an outlier for Z because the normal distribution is a location-scale family of distributions. Also, Z is symmetric about 0.

   $$Q(0.25) = -0.6745 \quad \Rightarrow \quad p_o = 2P(Z < 4Q(0.25)) = 2P(Z < 4(-0.6745)) \approx 2(0.0035) = 0.007.$$

5. **t with df=2** is symmetric about 0 and using R we have $Q(.25) = qt(.25, 2) = -.8164966\ \Rightarrow$

   $p_o = 2P(Y < 4Q(0.25)) = 2pt(4*(-.8164966), 2) = .0823$ alternatively $p_o = 2*pt(4*qt(.25, 2), 2) = .0823$

IV. (10 points) Let X be the number of breaks on a given bar. Then, the estimated probability of a break at a location on a randomly selected bar is given by

$$\hat{p} = \frac{(0)(121) + (1)(110) + (2)(38) + (3)(7) + (4)(3) + (5)(1)}{(5)(280)} = .16$$

Next evaluate if the distribution of X is Binomial$(5, p)$.

Under the Binomial model, $p_i = \binom{5}{i} p^i (1-p)^{5-i}$, for $i = 0, 1, 2, 3, 4, 5$.

Since p is unknown use $\hat{p} = .16$.

An initial calculation of $E_i = 280 p_i$ shows that both E_4 and E_5 are less than 1.

After combining the last two cells, the expected count is still less than 1.

Therefore, combine the last three cells and then compute the following using R-functions:

$p_i = dbinom(i, 5, .16)$ for $i = 0, 1, 2$ and $p_3 = P[X \geq 3] = 1 - P[X \leq 2] = 1 - pbinom(2, 5, .16)$,

finally compute $E_i = 280 * p_i$.

```
     i     pi          Ei        Oi   (Oi-Ei)^2/Ei
[1,] 0 0.41821194 117.099344    121   0.12993342
[2,] 1 0.39829709 111.523185    110   0.02080367
[3,] 2 0.15173222  42.485023     38   0.47347106
[4,] 3 0.03175875   8.892449     11   0.49949933
```

The test statistic is

$$Q^* = \sum_{i=1}^{4} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = .130 + .021 + .473 + .499 = 1.12$$

and Q^* has approximately a chi-squared distribution with df$=4 - 1 - 1 = 2$.

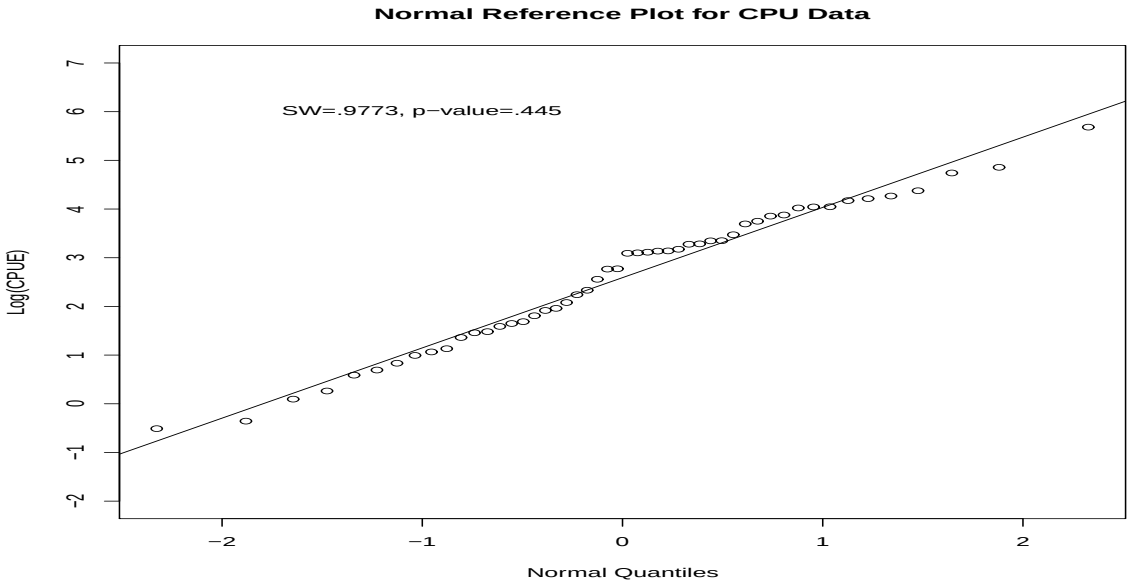
The p-value$= Pr(\chi^2_2 \geq 1.12) = 1 - pchisq(1.12, 2) = 0.571$.

Thus, we conclude that there is an excellent fit of the Binomial model to the data.

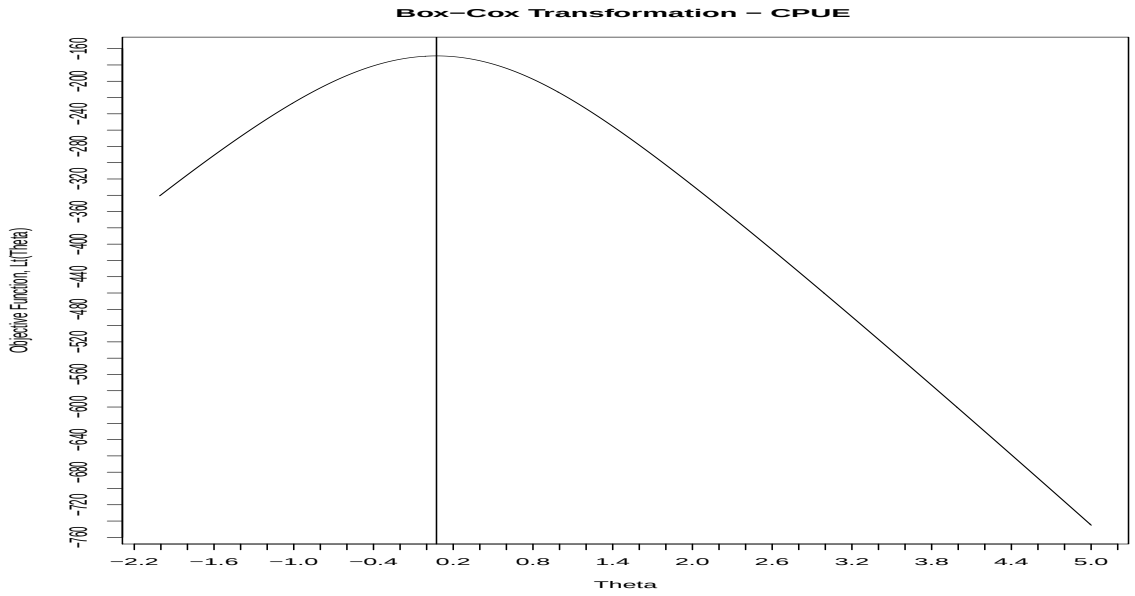
Notice that the expected counts for the four cells under the binomial model are very close to the observed counts and all the expected counts are greater than 5.

V. (10 points) CPUE problem.

1. The Normal Reference Distribution plot has nearly all the 50 close to a straight line. The Shapiro-Wilk test has the following values:
   SW=.9773, p-value = .445.
   Thus, we would conclude there is an excellent fit of the log normal model to the observed CPUE data.



2. Because the Log-Normal fit the CPUE data so well it is not necessary to perform a power transformation using the Box-Cox transformation. However, if you conduct the Box-Cox process on the CPUE data, we obtain $\widehat{\theta}_{max} = .074$ with a 95% CI of (-.116, .264) which contains 0, the log transformation. The Shapiro-Wilk test on $(CPUE)^{.074}$ yields SW=.9815 and p-value=.617 which is a slight improvement over the fit for the Log(CPUE) data.



VI. (20 points) Note: the models for the Weibull distribution are in their alternative form with $\beta = \alpha^{\gamma}$. For the 4 plots we have

Plot 1: J - because Plot 1 is a mixture distribution with $Q(.5) \approx 10$

Plot 2: E - because Plot 2 is a symmetric distribution with both tails lighter (shorter) than a normal distribution and all its mass on (0,.7)

Plot 3: I - because Plot 3 displays a right skewed distribution with $Q(.5) \approx 22$ and

Gamma(1.2,25) has $Q(.5) = 22.2 < \mu = \alpha\beta = 30$

Weibull(0.7,20) has $Q(.5) = [-20log(.5)]^{1/.7} = 42.8$

Exponential(80) has $Q(.5) = [-80log(.5)] = 55.45$

Plot 4: G or H - because Plot 4 displays a right skewed distribution with

$$\hat{Q}(.159) \approx 15, \hat{Q}(.5) \approx 45, \hat{Q}(.75) \approx 110, \hat{Q}(.84) \approx 150, \hat{Q}(.977) \approx 280$$

- Gamma(1.2,25) has $Q(.5) = 22.2 < \mu = \alpha\beta = 30$
- Weibull(0.7,20) has $QW(u) = [-20(1 - log(u)]^{1/.7} \Rightarrow$

  $QW(.159) = 6,\ QW(.5) = 43,\ QW(.75) = 115,\ QW(.84) = 172,\ QW(.977) = 481$
- Exponential(80) has $QE(u) = [-80log(1-u)] \Rightarrow$

  $QE(.159) = 13,\ QE(.5) = 55,\ QE(.75) = 111,\ QE(.84) = 147,\ QE(.977) = 302$

VII. (18 - 3 pts each) Multiple Choice Questions.

1. **D** - First sentence on page 16 in Handout 9.
2. **E** - If $F_o(y)$ does not have any unspecified parameters then use A-D but if $F_o(y)$ is normal then use Shapiro-Wilk, etc.
3. **D** - See page 17 in Handout 9.
4. **D** - See page 10 in Handout 9.
5. **C** - The distributions for Y are for a discrete random variable.
6. **D** - As the sample size increases, the sensitivity of the GOF test increases and hence small deviations in the data from the claimed model are enhanced.

IV. (10 points) Let X be the number of breaks on a given bar. Then, the average number of breaks per bar is

$$\bar{X} = \frac{(0)(121) + (1)(110) + (2)(38) + (3)(7) + (4)(3) + (5)(1)}{280} = .8$$

Next evaluate if the distribution of X is Poisson(λ). Under the Poisson model, $p_i = e^{-\lambda}\lambda^i/i!$, for $i = 0, 1, 2, 3, 4$ and $p_5 = P[X \geq 5]$. Since λ is unknown use $\hat{\lambda} = \bar{X} = .8$. An initial calculation of $E_i = 280p_i$ shows that E_5 is less than 1. Therefore, combine the last two cells and then compute the following using the R-function $pi = dpois(i, .8)$ for $i = 0, 1, 2, 3$ and $p4 = P[X \geq 4] = 1 - P[X \leq 3] = 1 - ppois(3, .8)$, finally compute $E_i = 280 * p_i$.

```
     i     pi          Ei        Oi    (Oi-Ei)^2/Ei
[1,] 0 0.449328964 125.81211    121       0.184
[2,] 1 0.359463171 100.64969    110       0.869
[3,] 2 0.143785269  40.25988     38       0.127
[4,] 3 0.038342738  10.73597      7       1.300
[5,] 4 0.009079858   2.54236      4       0.836
```

$$\begin{aligned}
\widehat{p}_0 = e^{-.8}(.8)^0/0! = \text{dpois}(0, .8) = 0.4493 \quad &\Rightarrow \quad \widehat{E}_0 = (280)(\widehat{p}_0) = 125.81 \\
\widehat{p}_1 = e^{-.8}(.8)^1/1! = \text{dpois}(1, .8) = 0.3595 \quad &\Rightarrow \quad \widehat{E}_1 = (280)(\widehat{p}_1) = 100.65 \\
\widehat{p}_2 = e^{-.8}(.8)^2/2! = \text{dpois}(2, .8) = 0.1438 \quad &\Rightarrow \quad \widehat{E}_2 = (280)(\widehat{p}_2) = 40.26 \\
\widehat{p}_3 = e^{-.8}(.8)^3/3! = \text{dpois}(3, .8) = 0.0383 \quad &\Rightarrow \quad \widehat{E}_3 = (280)(\widehat{p}_3) = 10.74 \\
\widehat{p}_4 = P[X \geq 4] = 1 - P[X \leq 3] = 1 - \text{ppois}(3, .8) = 0.0091 \quad &\Rightarrow \quad \widehat{E}_4 = (280)(\widehat{p}_4) = 2.54
\end{aligned}$$

The test statistic is

$$Q^* = \sum_{i=1}^{4} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = .184 + .869 + .127 + 1.300 + .836 = 3.316$$

and Q^* has approximately a chi-squared distribution with df=$5 - 1 - 1 = 3$.
The p-value= $Pr(\chi^2_3 \geq 3.316) = 1 - pchisq(3.316, 3) = 0.345$ and so we conclude that there is an excellent fit of the Poisson model to the data. Notice that the expected counts for the five cells under the Poisson model are very close to the observed counts and all but one expected count are greater than 5 with the remaining value greater than one.