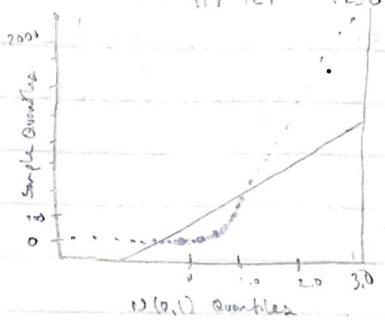
Jack Rodan
STAT 641
Assignment 5
10/25/21

See Handouts 8 & 9 & 10, Chp 1, 6, 14, sections 4.6, 5.3 - 5.6 & 7.2 - 7.4 in Devore book.

(H.O.4) 1) 250 iid observations $Y_1, Y_2, \dots Y_{250}$ from a process yield the following normal probability plot.




- Using the plot given, describe the process distribution relative to a normal plot w/ respect to tail weight and symmetry.

- From the Q-Q plot shown we can tell that the process distribution has a heavier right tail and lighter left tail than the $N(0,1)$ distribution. Furthermore, the process distribution is skewed heavily to the right.

- * need to redo because $\hat{E}_6 < 1$. Need to combine last two groups. ☑

(H.O.9) 2.) Nylon Bars were tested for brittleness. Each of the 500 bars was modeled under similar conditions and was tested by placing a specified stress at 5 locations on the bar. Assuming each bar has uniform composition, the number of breaks on a given bar should be binomially distributed w/ an unknown probability $p$ of breaking. The following table summarizes the outcome of the experiment:

| Breaks / Bar | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| Frequency | 140 | 197 | 115 | 41 | 5 | 2 | 500 |

- Use a GOF test to evaluate whether the data appears to be from a binomial model.

① Find MLE of $\theta$; $\hat{\theta}$ = Sample probability of a location on a bar breaking.

$$\hat{\theta} = \frac{0(140) + 1(197) + 2(115) + 3(41) + 4(5) + 5(2)}{500(5)} = \underline{0.232 = \hat{\theta}}$$

② Let $p_{i+1}$ be the probability that a randomly selected bar has $i$ breaks for $i \in [0, \dots, 5]$. That is $p_{(i+1)} = P[X = i]$ where $X \sim \text{Binomial}(5, \theta)$ distribution

- $\hat{p}_{i+1} = P(X = i) = f_0(i, \hat{\theta}) = \binom{5}{i} (0.232)^i (0.768)^{5-i}$ for $i \in [0, \dots 5]$
- $E_i = 500\, \hat{p}_i$
- $\hat{Q} = \sum_{i=1}^{5} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$

③ Following table was calculated using R.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| $\hat{p}_i$ | 0.26978813 | 0.43585231 | 0.2818814 | 0.07 3652 | 0.01112457 | 0.00067210 93 | 1.00 |
| $\hat{E}_i$ | 134.89066 | 201.79256 | 121.90727 | 36.82615 | 5.562284 | 0.3360546 6 | 500 |
| $O_i$ | 140 | 197 | 115 | 41 | 5 | 2 | 500 |
| $\hat{Q}_i = \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$ | 0.3075 | 0.1131 | 0.3913 | 0.4730 | 0.05684 | 8.238 88 | 9.58077 |

NOTE: $\hat{E}_6 < 1 \Rightarrow$ Need to recalculate w/ 5 and 6 combined.

2) (cont'd)

$Q_{new} = Q_{old} - E_{y,old} - E_{y,old} + E_{y,new} = 9.980779 - 5.542284 - 0.926054 + \frac{(7 - 5.899229)^2}{5.899229}$

$Q_{new} = 1.490813$ $\quad df_{new} = 5 - 1 - 1 = 3$

$P(Q \geq 1.490813) = 1 - P(Q \leq 1.490813) = 1 - p\text{chisq}(1.490813, 3)$

$P(Q \geq 1.490813) = 0.6844223$ $\Rightarrow$ w/ a p-value. but large, we would conclude a normal distribution is a good fit of our data.

(+0.8) 3.) A random sample of 500 data values are selected from 4 separate processes having cdfs $F_1, F_2, F_3, F_4$. The plot of the sample quantiles vs a standard normal quantile for each of the four samples is given below. For each of these plots, select one of the following distributions to describe the pdf which generated the data. (Hint: make sure to take into consideration the size of values associated w/ each distribution).

Plot 1: K

Plot 2: D (b/c $x_2$ goes from 0 to 1)

Plot 3: H

Plot 4: G

4.) An experiment was conducted to investigate if the impact of the carcinogen DMBA could be delayed by treatment w/ a potential beta-blocker. 50 mature rats of the same general health were given the beta-blocker then injected w/ DMBA. The time in days, after exposure, at which the carcinoma was diagnosed for the rats are given below. From past studies, 150 days after exposure, a carcinoma was detected in untreated rats.

*Data given in table*

- Does a weibull distribution appear to provide an adequate fit to the data? Justify your answer using both a GOF and a graphical plot.

① Use R to find MLE of $\gamma$ ; $\alpha$, $\hat{\gamma}$ ; $\hat{\alpha}$

$\hat{\gamma} = 1.3861041$, $\hat{\alpha} = 201.17866473$

AD = 0.24708605

ADM = 0.2232267

- From Table 5 in H.O. 9 (pg 34) we see that our p-value is $> 0.25$. Thus we have a very good fit.
- Similarly, looking at our QQ plot in R we see that the weibull model is a good fit.

*note corr test and AD are not matching up. AD => good fit, corr test => not good.

*Ask about corr test for non normal data.

$X^{\theta_{MAX}} \sim N$

→ might want to redo 0 using the shapiro-wilks test instead of AD

A major problem in the Gulf of Mexico is the excessive capture of game fish by shrimpers. A random sample of the catch of 50 shrimpers yield the following data concerning the catch per unit effort (CPUE) of Red Snapper, a highly sought game fish. Let $C_i$ be the CPUE of the $i^{th}$ shrimper. The data $C_1, \dots, C_{50}$ is given in the below table.

(1.) CPUE data is often modeled using a log-normal distribution. Does the above data appear to be from a Log-Normal distribution? Explain your answer w/ both a normal reference distribution plot and a GOF test.

- Let $X_i = \ln(C_i)$.
  Then $X_i \sim N(\mu, \sigma^2)$

Make it clear this is shapiro-wilks instead of AD.

- Using MLE estimates from R, we get $\hat{\mu} = 2.590627$, $\hat{\sigma}^2 = 2.070229$
  AD = 0.4774204, ADM = 0.4550114.
- From table 5 (4.0.9 pg 34) we see that our p-value is btwn (0.10, 0.25) thus the lognormal distribution provides a moderately good fit of the data

- In R: ldata = log(data), x = sort(ldata). shapiro.test(x)
  W = 0.97731, p-value = 0.4451 => the lognormal distribution provides an "excellent fit" (4.0.9 pg 5) to our data

(2.) Use the Box-Cox transformation of the CPUE data to determine the most appropriate power transformation to transform the CPUE distribution to Normality. How does the fit from the box-cox transformation compare to the fit for the log transformation?

- From R: $\hat{\theta} = 0.074$;
- In R: data = data^0.074 ; shapiro.test(data)
  => W = 0.98151, p-value = 0.6173.
  The fit from the boxcox transformation is better that the fit for the log transformation

5.) (Contd)

(3) Use the R program from H.O. 10 to draw 10,000 Bootstrap samples from the CPUE data. From the 10,000 samples, estimate the standard error of the sample mean for the $Y_i = \log(C_i)$ data. Compare this estimate to the usual estimate $S_Y/\sqrt{n}$, where $S_Y$ is the sample standard deviation computed from the $n=50$ values of $Y_i = \log(C_i)$.

$\frac{S_Y}{\sqrt{n}} = 0.20554664$

$SD(\bar{Y}) = 0.20357657$

The estimated standard error of the sample mean is approximately equal to the sample standard error.

(4) Use your bootstrap samples to estimate the mean and standard deviation of the following sample statistics for $Y_i = \log(C_i)$

(a) sample median; $\hat{Q}(0.5)$

$\bar{\hat{Q}}(0.5) = 2.8149869$, $SD(\bar{\hat{Q}}(0.5)) = 0.3721707$

(b) sample SD; $S_Y$

$\bar{S}_Y = 1.433990$, $SD(\bar{S}_Y) = 0.1187079$

(c) the sample MAD; $\hat{MAD}_Y$

$\overline{MAD} = 1.5698568$, $SD(\overline{MAD}) = 0.2244567$

(H.O.10 pg 18) (5) Historically, the log(CPUE) data was modeled as a random sample from a $N(3, (1.5)^2)$ distribution. Compare your bootstrap estimates of the mean and standard deviation of $\hat{Q}(0.5)$ and $S_Y$ from part 4 of the problem to the theoretical mean and standard deviation of $\hat{Q}_Y(0.5)$ and $S_Y$ based on $Y_i = \log(C_i)$ having a $N(3, (1.5)^2)$ distribution.

- Asymptotic Mean of $\hat{Q}(0.5)$ is $\mu = Q(0.5) = 3$
  - This is approximately the same as the mean of the sample median
- Asymptotic SD of $\hat{Q}(0.5)$: $\sigma_A = \frac{\sqrt{0.5 \cdot 0.5}}{\sqrt{50}\, f(Q(0.5))} = \frac{\sqrt{0.5^2}}{\sqrt{50}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} = \frac{1.5\sqrt{\pi/2}}{\sqrt{50}} = 0.26586681$
  - The asymptotic SD of $\hat{Q}(0.5)$ is less than our estimated SD of $\hat{Q}(0.5)$.
- Asymptotic Mean of $S_Y$ is $\sigma = 1.5$
  - The asymptotic mean of $S_Y$ is approximately equal to our estimate $\bar{S}_Y$.
- Asymptotic SD of $S_Y$ is $\sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}}$, $\mu_4 = 3\sigma^4 = 3(1.5)^4 = 15.1875$

$\sigma_A = \sqrt{15.1875 - (1.5)^4} \,/\, (2(1.5)\sqrt{50}) = \underline{0.15}$

- The Asymptotic SD of $S_Y$ is slightly larger than our estimate.

$SD(\sigma_Y)$ -

6.) A company has designed a new battery system for electric powered automobiles. To estimate the lifetime of the systems, the design engineers place the batteries in 25 electric powered cars and test them under simulated city driving. Let $Y_i$ be the time to failure of the batteries of the $i^{th}$ car, $i \in [1, \ldots, 25]$. The failure times are recorded in units of 20,000 miles. The company wants to know the probability that the sample mean based on 25 observations will estimate the true mean w/in a margin of error of $\pm 0.2$ (40000 miles), provided that the true mean has a value of 5 (100,000 miles), that is, approximately, $P[-0.2 \le \bar{Y} - 5 \le 0.2]$
$= P[4.8 \le \bar{Y} \le 5.2]$

(1). From past studies, the distribution of the time to failure of the braking system is exponential w/ a mean value of 5 (100000 miles). Let $\bar{Y}$ be the sample mean time to failure of the 25 cars.

(4.0 pts) (a) What is the exact distribution of $\bar{Y}$ if the exponential model is still valid?

[*In thinking we use the fact that if $X_1, \ldots, X_n$ i.i.d. $Exp(\beta)$, then the distribution $n\hat{\beta} = \sum_{i=1}^{n} X_i \sim Gamma(n, \beta)$]

- Let $X = \sum_{i=1}^{25} Y_i \Rightarrow X \sim Gamma(25, 5)$

Then $\bar{y} = \left(\frac{1}{25}\right)X \Rightarrow X = h^{-1}(\bar{y}) = 25\bar{y}$ $\quad \frac{d}{d\bar{y}}\left[h^{-1}(\bar{y})\right] = 25$

$$f_{\bar{Y}}(\bar{y}) = \frac{1}{\Gamma(25)\,5^{25}} (25\bar{y})^{25-1} e^{-(25\bar{y}/5)} (25)$$

$$= \frac{5^{50}}{\Gamma(25)\,5^{25}} \bar{y}^{25-1} e^{-(5\bar{y})} = \frac{5^{25}}{\Gamma(25)} \bar{y}^{25-1} e^{-5\bar{y}}$$

$$= \frac{1}{\Gamma(25)\left(\frac{1}{5}\right)^{25}} \bar{y}^{25-1} e^{-5\bar{y}} \Rightarrow \boxed{\bar{Y} \sim Gamma\left(25, \tfrac{1}{5}\right)}$$

(c) What is the mean and SD of $\bar{Y}$ if the exponential model is still valid?

$$E[\bar{Y}] = 25\left(\tfrac{1}{5}\right) = \boxed{5 = E[\bar{Y}]}$$
$$SD(\bar{Y}) = \sqrt{25\left(\tfrac{1}{5}\right)^2} = \boxed{1 = SD(\bar{Y})}$$

(2) Simulate 10000 random samples of size 25 from the exponential distribution w/ $\beta = 5$. Compute the sample mean for each of the 10000 samples. Display a normal reference plot for the 10000 sample means. Does the plot suggest that the sampling distribution of $\bar{Y}$ is approximately normal?

• NO, the sampling distribution of $\bar{Y}$ is not approximately normal

((ontd)

(3) Compute or estimate $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$ in each of the following ways

(a) Using the exact distribution of $\bar{Y}$

- $P[-0.2 \leq \bar{Y} - 5 \leq 0.2] = P[4.8 \leq \bar{Y} \leq 5.2]$

$= P[\bar{Y} \leq 5.2] - P[\bar{Y} \leq 4.8]$

$P[-0.2 \leq \bar{Y} - 5 \leq 0.2] = 0.1580747$

(b) Using Central Limit Theorem

0.1585194

(c) Using your simulated 10000 $\bar{Y}$'s

$\hat{P}[-0.2 \leq \bar{Y} - 5 \leq 0.2] = 0.1546$

(4) What is the level of agreement in the three computations/estimates of $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$?

- All three estimates agree to the nearest hundredth, the central Limit Theorem and the exact distribution agree to the nearest thousandth.