Q1

(a) The objective of the study is to determine if the drug, Lasix, should be given to all horses regularly

(b) The units are the horses

(c) The response variate is the effect of the drug on the horse, for example, if the horses bleed, or if their results increased compared to previous races without the drug, etc.

(d) the Explanatory Variate is the horses that have taken the drug or not (i.e let x represent if the horses have (1) or don't have the drug (0))

(e) The attribute could be the average response of the drug, or the average affect the drug has on the horses during the race. Another could be the proportion of horses who placed in the top placements during the race who are on the drug compared to those who aren't.

(f) The study population are the 100 horses apart of the study (which will represent the total population of all the horses).

(g) Since we're only testing on race horses, a source of study error could be that we don't know it's effects on regular or work horses. Another source of study error could be that the drug has minimal effects and we are unable to make any conclusions regarding the effectiveness of the drug.

(h) Some horses could potentially be ill, or have some problems which complicate the affect of the drug on the horse. Also trivially randomness could make rise for some sampling error.

(i) Unable to clearly make out a relationship of the drug and the performance of the horse during the race due to its minimal effects. Another could be that due to the rough and dirty terrain, the tester are unable to discern whether or not bleeding is present unless the horse is clean and all parts of body are visible.

Q2

Below is the table of bins and frequencies used in the problem.

Bin	Frequency
$[0, 20)$	114
$[20, 40)$	64
$[40, 60)$	74
$[60, 80)$	33
$[80, +\infty)$	35

(a) We'll need some numbers before we begin. We have that the sample mean and the rate parameter are

$$\hat{\mu} = \frac{1}{320} = \frac{1}{320}\sum_{i=1}^{320} y_i \approx 35.294$$
$$\lambda = \frac{1}{\hat{\mu}} \approx 0.0283$$

We know that the c.d.f of the exponential is $F(x) = 1 - e^{-\lambda x}$, and thus, the expected counts for each bin is simply

$$e_1 = n \cdot P_1 = 320 \cdot P(0 \leq x \leq 20) = 320(F(20) - F(0)) \approx 138.429$$
$$e_2 = n \cdot P_2 = 320 \cdot P(20 \leq x \leq 40) = 320(F(40) - F(20)) \approx 78.546$$
$$e_3 = n \cdot P_3 = 320 \cdot P(40 \leq x \leq 60) = 320(F(60) - F(40)) \approx 44.568$$
$$e_4 = n \cdot P_4 = 320 \cdot P(60 \leq x \leq 80) = 320(F(80) - F(60)) \approx 25.288$$
$$e_5 = n \cdot P_5 = 320 \cdot P(80 \leq x \leq \infty) = 320(F(\infty) - F(80)) \approx 33.169$$

With this, we can plot this in the same graph and compare the values From this it's fairly obvious that

Bin	Frequency	Expected Frequency
$[0, 20)$	114	138.492
$[20, 40)$	64	78.546
$[40, 60)$	74	44.568
$[60, 80)$	33	25.288
$[80, +\infty)$	35	33.169
total	320	320

the expected frequency (which uses an exponential distribution) does not follow the frequencies observed. The first two are much higher, the third is much much lower, and the last two are again under.

(b) I guess that this data will still not fit the exponential distribution since, although the bins have changed, the data is still the same and the previous part showed that it would not follow the exponential distribution. Since it's the same data set, we have the same sample mean and the same rate parameter,

therefore we have that

$$
\begin{aligned}
e_1 &= n \cdot P_1 = 320 \cdot P(0 \leq x \leq 5) = 320(F(5) - F(0)) \approx 42.269 \\
e_2 &= n \cdot P_2 = 320 \cdot P(5 \leq x \leq 10) = 320(F(10) - F(5)) \approx 36.696 \\
e_3 &= n \cdot P_3 = 320 \cdot P(10 \leq x \leq 20) = 320(F(20) - F(10)) \approx 59.474 \\
e_4 &= n \cdot P_4 = 320 \cdot P(20 \leq x \leq 40) = 320(F(40) - F(20)) \approx 78.546 \\
e_5 &= n \cdot P_5 = 320 \cdot P(40 \leq x \leq 80) = 320(F(80) - F(40)) \approx 69.856 \\
e_6 &= n \cdot P_5 = 320 \cdot P(80 \leq x \leq \infty) = 320(F(\infty) - F(80)) \approx 33.169
\end{aligned}
$$

Comparing this in a table as we did in part (a) shows us that Which, as we guessed above, we notice

Bin	Frequency	Expected Frequency
$[0, 5)$	4	42.269
$[5, 10)$	40	36.696
$[10, 20)$	70	59.474
$[20, 40)$	64	78.546
$[40, 80)$	107	69.856
$[80, +\infty)$	35	33.169
total	320	320

that the exponential distribution still doesn't fit.

Q3

(a) Since we know that the Gamma distribution is the sum of k i.i.d values from an exponential, then we know that (for $Y \sim Gamma(k, \theta)$, and $X \sim Exponential(\theta)$), we have that

$$Y = \sum_k X_i \implies E[Y] = E\left[\sum_k X_i\right] = \sum_k E[X_i] = k\theta$$
$$\implies \text{Var}[Y] = \text{Var}\left[\sum_k X_i\right] = \sum_k \text{Var}[X_i] = k\theta^2$$

With this, we have that

$$E\left[\frac{Y}{k}\right] = \frac{1}{k}E[Y] = \theta$$
$$\text{Var}\left[\frac{Y}{k}\right] = \frac{1}{k^2}\text{Var}[Y] = \frac{\theta^2}{k}$$

(b) CLT stats that as the number of samples, k, grows sufficiently large, the distribution of the sample mean approaches a normal distribution. In our case, the distribution of the sample mean is simply

$$\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$$

and by CLT, will approach a normal distribution where $Y \sim N(\theta, \theta/k^2)$ (where we found both the mean and the variance above). It also helps to note that both the mean and variance of Y are finite since they come from the exponential distribution.

(c) We can use the CLT to our advantage and approximate this distribution as a normal distribution. Since we have that $\theta = 5$, we'd have that $Y/k \sim N(5, 25/k)$. With this, we can standardize the probability as follows

$$P\left(\theta - 0.3 \leq \frac{Y}{k} \leq \theta + 0.3\right) \to P\left(\frac{\theta - 0.3 + \theta}{\sqrt{25/k}} \leq \frac{Y/k - \theta}{\sqrt{25/k}} \leq \frac{\theta + 0.3 + \theta}{\sqrt{25/k}}\right)$$

We can simplify the intervals and find that we're simply looking for

$$P\left(|Z| \leq \frac{0.3}{\sqrt{25/k}}\right) = 0.95$$

Looking at a normal table, we have that the value which corresponds to this probability is approximately 1.96. Therefore, we have that

$$\frac{0.3}{\sqrt{25/k}} = 1.96 \implies k \approx 1067.1$$

and we round up to $k = 1068$ since we would rather be certain it is included.

Q4

We know from class that we can construct a 95% confidence interval by the following $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. So, for each of these, we simply find $\hat{\theta}$ and investigate the confidence interval, let us make a table and stop when the confidence interval does not contain 0.1

Sample Size	$\hat{\theta}$	CI
50	8/50	$[0.058, 0.261]$
100	17/100	$[0.096, 0.243]$
150	21/150	$[0.084, 0.196]$
200	32/200	$[0.109, 0.211]$

There we go! When we have a sample size of 200, the confidence interval doesn't contain 0.1 anymore! The interval has exceeded this value. From this, we conclude that the proportion of defective to non defective products is greater than 0.1

Q5

(a) See attached image

(b) The 5% likelihood interval seems to be $[37.7, 42.2]$. The 10% likelihood interval is between $[38.1, 42]$. And finally, the 20% likelihood interval seems to be between $[38.5, 41.5]$

(c) I noticed that the likelihood intervals get smaller the higher percent we are looking at.

(d) With this increases sample size, the 10% likelihood interval becomes somewhere around $[39.6, 40.4]$

(e) As the sample size increases, the intervals become smaller and smaller (more and more accurate from the increased sampling).