BEE552 Biometry Midterm Rewrite

Maria Feiler

03/29/2022

My Learning Journey

Over the last week, I participated in Biometry in the following ways:

- I asked / answered **4** questions posed in class.
- I asked **0** questions in Slack.
- I answered **0** questions posed by other students on Slack.
- I came to Heather's office hours: **No**
- I came to Jose's office hours: **No**
- I met with Heather or Jose separately from office hours: **No**

Anything not falling into one of the above categories?

> **No**

On a scale of 1 (no knowledge) to 10 (complete expert), how would I rate my comfort with R programming after this week?

> **7**

Any topics from last week that you are still confused about?

> **Will message in Slack**

Maria Feiler
11435783Z

Biometry mid-term exam Spring 2022

(200 pts total)

Section 1 – Short answer

(I would spend no more than 45 minutes maximum on the short answer section in order to leave yourself enough time for the long answer problems.)

1. (20 pts total; 10 pts each) (True story: Liliana Davalos needs our help! She is trying to model the data represented in the histogram below.) Below is a histogram of the number of years (recorded as an integer) between the start of a study until a patch of forest is cleared. The bins are inclusive of the number on the right of the interval (in other words, the first bin is the number of patches with Time=0 years, the second bin represents Time=1 year, the third bin represents Time=2 years, etc.)

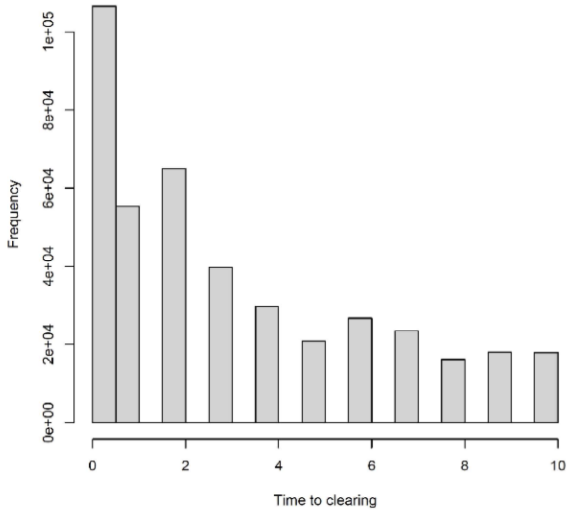


Figure 1: Number of years (integer) until a patch of forest is cleared.

a) Name one distribution (including ballpark estimates of the distributions parameter(s)) that could have generated this dataset.

Poisson distribution because discrete data that's zero-bounded

$P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\lim_{x\to\infty} \text{Poisson}(\lambda) \to N(\lambda,\lambda)$ $\lambda \approx 3$, $\therefore$ $E[X] = \lambda$

$sd \approx 2$ $\text{Var}[X] = \lambda$

b) Name a second distribution (including estimates of the distributions parameter(s)), different from the one identified in part a, that could be used to model these data. Keep in mind that sometimes we use a distribution that know couldn't be the generating distribution but is "good enough" to use for modelling the data.

Gamma distribution because though not continuous data, the data is right skewed & zero bounded

partial credit, did not provide estimates for α & β

$f(x|\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{(\alpha-1)} e^{-x/\beta}$ $\alpha\beta \approx \lambda \approx 3$

$\lim_{x\to\infty} \text{Gamma}(\alpha,\beta) \to N(\alpha\beta, \alpha\beta^2)$ $sd \approx 2$

$\text{Var} = sd^2 = 4$

$E[X] = \alpha\beta \approx 3$

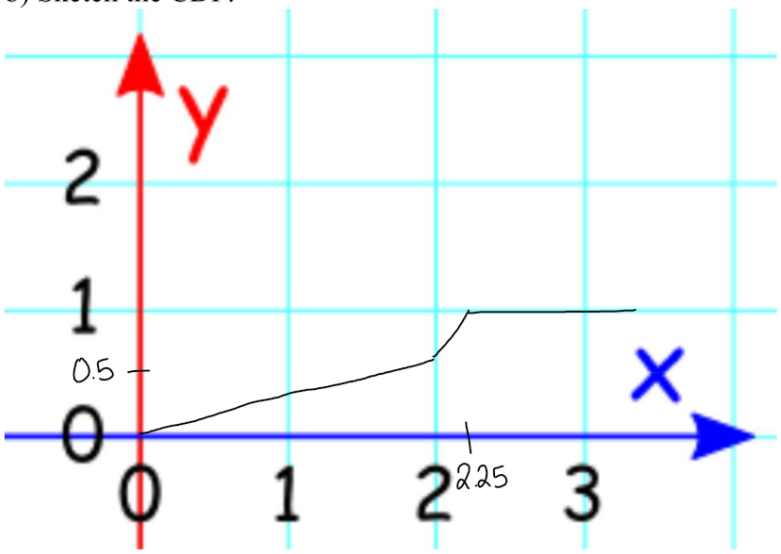
$\text{Var}[X] = \alpha\beta^2 \approx 4$

2. (20 pts total; 5 pts each)
Consider the following PDF:

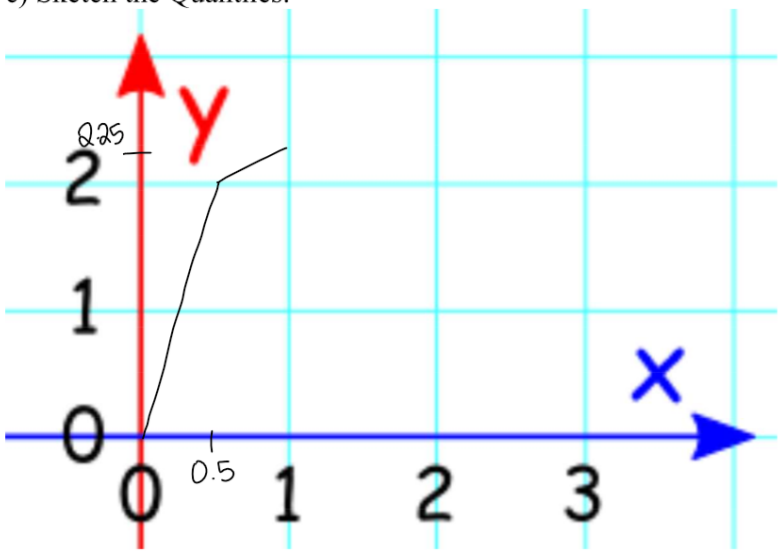
$$f(x) = \begin{cases} 0.25 \; for \; 0 \leq x < 2 \\ 2.0 \; for \; 2 \leq x < 2.25 \\ 0 \; otherwise \end{cases}$$

a) Sketch the PDF. Full credit

b) Sketch the CDF. Full credit



c) Sketch the Quantiles.



d) Provide a possible set of 5 random draws from this distribution. full credit

1, 2.1, 2.15, 2.2, 2.21

more likely because the PDF (f(x)) is higher from 2 to 2.25*

* all the same as original correct answer except better defined the range

3. (10 pts total; 2.5 pts each) Consider the following Venn diagram from Foley* et al. (2017) summarizing the number of species falling into each of the three protection regimes: 1) Being listed in Appendix I of the CITES treaty, 2) Being listed as IUCN Threatened, or 3) Being listed under the Endangered Species Act. In this question, we considered only the 1509 species depicted in this diagram. (*Foley was a former Ecology & Evolution Ph.D. student and a Biometry alumna.)

a) What is the P(CITES listed, IUCN Threatened)?

not independent

$P = \frac{33 + 67}{1509} = \frac{100}{1509}$

b) What is the P(ESA listed | CITES listed)?

$\frac{P(ESA \cap CITES)}{P(CITES)} = \frac{\frac{102}{1509}}{\frac{155}{1509}} = \frac{102}{155}$

c) What is the P(CITES listed or ESA listed)?

$P(CITES) + P(ESA) - P(CITES \cap ESA)$

$\frac{155}{1509} + \frac{233}{1509} - \frac{102}{1509} = \frac{256}{1509}$

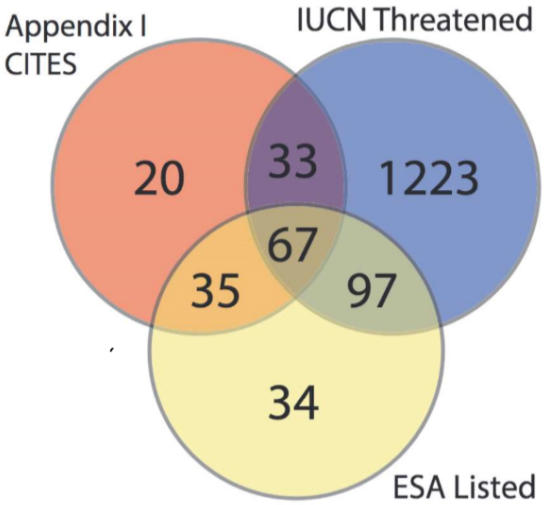


Figure 1. Venn diagram demonstrating the overlap between foreign bird species listed under the Endangered Species Act (ESA), foreign bird species listed under Appendix I of CITES, and foreign bird species in a threatened category as assessed by the IUCN Red List. Of the 34 listings unique to the ESA, 18 are subspecies of species not on CITES appendix I or deemed threatened by the IUCN.

d) What is the marginal probability of being IUCN Threatened?

$P(IUCN) = \frac{1420}{1509}$

35 + 67 = 102

97 + 67 + 35 + 34 = 233

1223 + 33 + 67 + 97 = 1420

20 + 33 + 67 + 55 = 155

4. (10 pts) The following is a list of confidence intervals for the parameters μ, σ^2, and ${\sigma_A}^2/{\sigma_B}^2$. Circle all correct expressions. If the expression is correct only in specific circumstances, explain. Assume that the degrees of freedom represented by "[dof]" are correct.

$$P\left(\frac{ns^2}{\chi^2_{(1-\alpha/2)[n]}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{(\alpha/2)[n]}}\right) = 1-\alpha$$

if σ unknown, μ known, finding σ

$$P\left(\bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(\alpha/2)} \le \mu \le \bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(1-\alpha/2)}\right) = 1-\alpha$$

if μ unknown, σ known, finding μ

$$P\left(\bar{X} - \sqrt{\frac{s^2}{n}} t_{(1-\alpha/2)[dof]} \le \mu \le \bar{X} + \sqrt{\frac{s^2}{n}} t_{(1-\alpha/2)[dof]}\right) = 1-\alpha$$

if μ & σ unknown, finding μ

$$P\left(\bar{X} \text{ (wrong: } -\text{)} \sqrt{\frac{\sigma^2}{n}} z_{(\alpha/2)} \le \mu \le \bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(1-\alpha/2)}\right) = 1-\alpha$$

equivalent

$$P\left(\frac{s_A^2}{s_B^2} F_{(\alpha/2)[m-1,n-1]} \le \frac{\sigma_A^2}{\sigma_B^2} \le \frac{s_A^2}{s_B^2} F_{(1-\alpha/2)[m-1,n-1]}\right) = 1-\alpha$$

$\frac{\sigma_A^2}{\sigma_B^2}$ unknown, $\therefore$ σ_A^2 & σ_B^2 unknown, finding ${\sigma_A^2}/{\sigma_B^2}$

$$P\left(\bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(1-\alpha/2)} \le \mu \le \bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(\alpha/2)}\right) = 1-\alpha$$

wrong

$$P\left(\bar{X} \text{ (wrong: } -\text{)} \sqrt{\frac{s^2}{n}} t_{(\alpha/2)[dof]} \le \mu \le \bar{X} + \sqrt{\frac{s^2}{n}} t_{(1-\alpha/2)[dof]}\right) = 1-\alpha$$

$$P\left(\bar{X} - \sqrt{\frac{\sigma^2}{n}} z_{(1-\alpha/2)} \le \mu \le \bar{X} + \sqrt{\frac{\sigma^2}{n}} z_{(1-\alpha/2)}\right) = 1-\alpha$$

if μ unknown, σ known, finding μ

$$P\left(\bar{X} - \sqrt{\frac{s^2}{n}} z_{(1-\alpha/2)} \le \mu \le \bar{X} + \sqrt{\frac{s^2}{n}} z_{(1-\alpha/2)}\right) = 1-\alpha$$

wrong

$$P\left(\frac{ns^2}{\chi^2_{(\alpha/2)[n-1]}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{(1-\alpha/2)[n-1]}}\right) = 1-\alpha$$

wrong

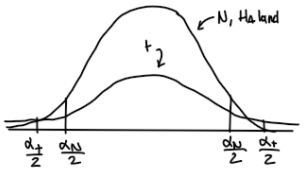
Biometry mid-term exam Spring 2022

5. (10 pts total; 5 pts each)

a) Define Type I error. full credit

Type I error is rejecting the null hypothesis even though the null is true, or the probability of rejecting the null hypothesis erroneously (set by critical value) added.

b) If a researcher accidentally used a Normal distribution as the distribution under the null hypothesis when the t-distribution was more appropriate, would the actual Type I error rate be higher, lower, or unchanged relative to the nominal (i.e. the intended) Type I error rate and why? full credit



Since the t is more conservative because of its wide tails, more probability sits in the tails. Therefore, the actual Type I error would be higher, pushing the confidence intervals wider & making it more likely the null is rejected.

6. (10 pts) When conducting an analysis involving multiple hypothesis tests, we worry about inflated Type I error rates. What happens to Type II error rates when we conduct multiple comparisons? (Full credit requires deriving a mathematical expression for the "family-wise" Type II error rate across *k* independent hypothesis tests.)

Type I and Type II error rates would inflate/increase towards 1 with multiple comparisons. Family-wise Type II error would be the product of all k probabilities of not making a Type II error where k is the number of comparisons. This is derived from the family wise Type I error rate.

$\alpha = 1-(1-\alpha')^k$ α = family wise Type I error rate, α' = per-comparison error rate

if β = family wise Type II error rate & β' = per-comparison error rate

then $\beta = 1-(1-\beta')^k$

Section 2 – Long answer

7. (40 pts) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables drawn from the Erlang distribution, whose probability density function is given by

$$f(x|k,\lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}, x \geq 0$$

a) (20 pts) Find the maximum likelihood estimator for the parameter λ.

1) joint likelihood
2) log likelihood
3) negative log likelihood
4) partial derivative, set to 0
5) solve for parameter

$$f(X \mid k,\lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

$$f(X \mid k,\lambda) = \prod_{i=1}^{n} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

$$L(k,\lambda \mid X) = \prod_{i=1}^{n} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

$$LL(k,\lambda|X) = \sum_{i=1}^{n} \left[\log(\lambda^k) + \log(x^{k-1}) + \log(e^{-\lambda x_i}) - \log(k-1)!\right]$$

$$= \sum_{i=1}^{n} \left[k\log\lambda + (k-1)\log x - \lambda x_i - \log(k-1)!\right]$$

$$\frac{\partial}{\partial\lambda}(NLL(k,\lambda|X)) = \left(\sum_{i=1}^{n}\left[\underbrace{\lambda x_i}_{=x_i} + \underbrace{\log(k-1)!}_{=0} - \underbrace{k\log\lambda}_{=k/\lambda} + \underbrace{(k-1)\log x}_{=0}\right]\right)\partial\lambda$$

correct until here

$$\frac{\partial NLL}{\partial\lambda} = \sum_{i=1}^{n}\left[x_i - \frac{k}{\lambda}\right] = 0$$

$$\sum_{i=1}^{n}(x_i) - \frac{kn}{\hat{\lambda}} = 0$$

$$\frac{\hat{\lambda}}{\sum x_i}\left(\sum_{i=1}^{n} x_i\right) = \frac{kn}{\hat{\lambda}}\left(\frac{\hat{\lambda}}{\sum x_i}\right)$$

$$\hat{\lambda} = \frac{kn}{\sum_{i=1}^{n} x_i}$$

b) (10 pts) How would you use bootstrap to find the standard error of the maximum likelihood estimator derived in part A?

*fixed from last time

1) Sample from $x_1, x_2, \ldots, x_n$ with replacement, with sample size n
2) calculate the parameter $(\hat{\lambda})$ for each sample
3) repeat many times
4) use the standard deviation* of the samples' $\hat{\lambda}$ values to calculate the standard error of the MLE

$$se_{boot} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\lambda}_i - \bar{\lambda})^2}$$

$\hat{\lambda}_i$ = i^{th} sample parameter estimate

$\bar{\lambda}$ = average value of parameter estimates

c) (10 pts) A special case of the Erlang function comes about when we set $k = 1$, as the Erlang distribution becomes what is known as the Exponential distribution. With $k = 1$, find E[X].

$$f(x|k,\lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \quad \text{if } k = 1 \ldots$$

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$= \int_{-\infty}^{\infty} x \lambda e^{-\lambda x}\,dx$$

$$= \lambda \left[uv - \int v\,du \right]_0^{\infty}$$

$uv - \int v\,du$

$u = x \qquad du = dx$

$dv = e^{-\lambda x}dx$

$v = \int dv = \int e^{-\lambda x}dx = \frac{e^{-\lambda x}}{-\lambda}$

$$= \lambda \left[x \cdot \frac{e^{-\lambda x}}{-\lambda} - \int \frac{e^{-\lambda x}}{-\lambda}\,dx \right]_0^{\infty}$$

$$= \lambda \left[-\frac{e^{-\lambda x} x}{\lambda} + \frac{1}{\lambda} \int e^{-\lambda x}\,dx \right]_0^{\infty}$$

$$= \left[e^{-\lambda x} x + \int e^{-\lambda x}\,dx \right]_0^{\infty}$$

$$= \left[-e^{-\lambda x} x - \frac{e^{-\lambda x}}{\lambda} \right]_0^{\infty}$$

$$= \left[-e^{-\lambda x} x \right]_0^{\infty} - \left[\frac{e^{-\lambda x}}{\lambda} \right]_0^{\infty}$$

$$= (0 - 0) - \left(0 - \frac{1}{\lambda} \right)$$

$$E[X] = \frac{1}{\lambda}$$

8. (40 pts)

As part of a study into the diets of the eastern horned lizard (*Phrynosoma douglassi brevirostre*), Powell and Russell (1984,1985) investigated whether the consumption of ants varied over time. The figure below contains boxplots of dry biomass of ants collected from the stomachs of 24 adult male lizards captured in the months June-September of 1980 (24 different lizards captured in each month). *In this question, we will restrict our attention to the months of June and July only.*

$\bar{X}$ = average biomass of ants

A = June B = July

a) (5 pts) What is the null hypothesis H_0 being tested? (Reminder: We are restricting our attention to June and July in this question.)

$H_0: \mu_A = \mu_B$

or $\mu_A - \mu_B = 0$

corrected $\bar{X}_A$ & $\bar{X}_B$ from last time

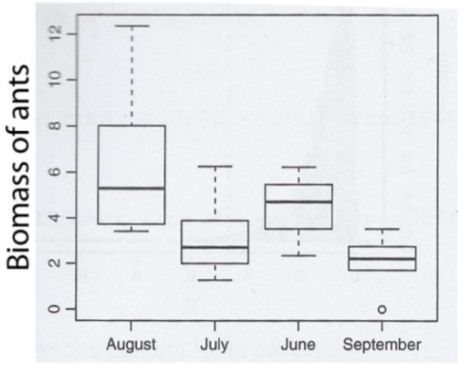


Figure 2: Boxplots of lizard consumption data.

b) (5 pts) What is the *best* parametric statistical test Powell and Russell should use to test this null hypothesis? (Here I am just looking for the full name of the test, not the equation.)

Full credit

unpaired two sample t-test assuming equal variance to increase statistical power since this experiment has very low sample sizes

c) (15 pts) Write the equation for this statistical test. (Remember that a complete description of the statistical test includes the test statistic and its distribution under the null hypothesis.) Define all variables in your expression and calculate the degrees of freedom for any distributions used.

$$T = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \qquad T|H_0 \sim t_{dof} \qquad dof = n_A + n_B - 2 = 46$$

as defined on previous page

- $\bar{X}_A$ = average biomass of consumed ants in June (estimate)
- $\bar{X}_B$ = average biomass of consumed ants in July (estimate)
- n_A = number of lizards captured in June = 24
- n_B = number of lizards captured in July = 24
- S_A^2 = estimate of variance in ant biomass in June
- S_B^2 = estimate of variance in ant biomass in July
- A = June
- B = July

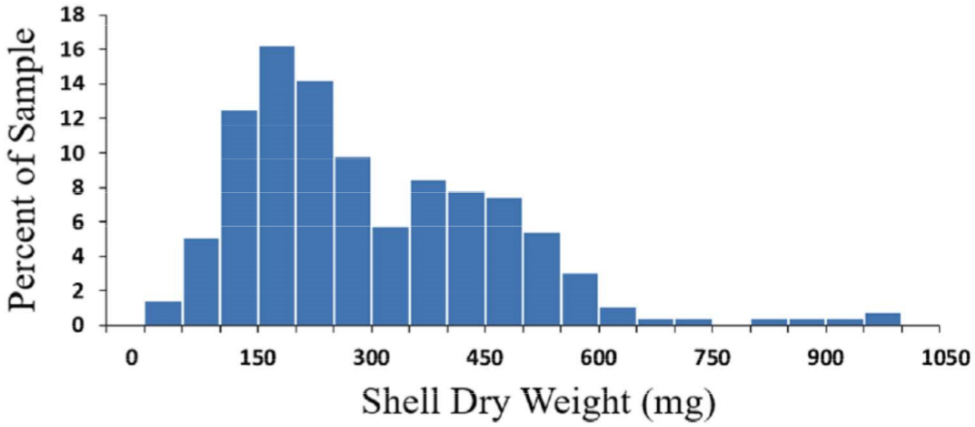
d) (15 pts) How might Powell and Russell have designed this study differently to increase the statistical power of their test without increasing their sample sizes? Explain the new experimental design and mathematically describe how/why this would increase the power of the test.

full credit * added

IF they had adjusted their catch & release technique such that they were capturing the same 24 lizards in June & July (tagging perhaps?*), then you could adjust for the potential variation between lizard ant consumption due to body mass/ sex/ health/ other factors. By using the same 24 lizards, you could perform a paired sample t-test which has higher power than an unpaired two-sample t-test

$$H_0: \bar{X}_{Ai} = \bar{X}_{Bi} \qquad T = \frac{\bar{\omega}_n}{\sqrt{\frac{S^2}{n}}} \quad \text{where } \omega_i = X_{A_i} - X_{B_i}$$

9. (40 points) Chase et al. (2020) studied the distribution of shell sizes among hermit crabs (*Pagurus longicarpus*) collected at West Meadow Beach. Hermit crab shell dry weight was recorded as a continuous variable, and was distributed as indicated in the histogram below.



Assume that Chase et al. (2020) modelled that data with a F distribution

$$X \sim F_{\alpha,\beta}$$

where I have labeled the two parameters of the F distribution as "α" and "β". We can use maximum likelihood to estimate the parameters of this F distribution $\hat{\alpha}$ and $\hat{\beta}$. We have two ways to express our uncertainty regarding our estimates of $\hat{\alpha}$ and $\hat{\beta}$: standard errors and confidence intervals.

a) (10 points) Describe in words the interpretation of the standard error of a parameter estimate ($\hat{\alpha}$ or $\hat{\beta}$) [Hint: The answer I'm looking for has something to do with *repeating* the experiment.]

After repeating the experiment many times & calculating the parameters (α & β) for each sample, the standard error is representative of your uncertainty of the estimate of the parameters for any given sample.

Standard error is also the standard deviation of the sampling distribution of the statistic.

b) (10 pts) Describe in words the interpretation of the confidence interval associated with the parameter estimates ($\hat{\alpha}$ or $\hat{\beta}$). [Hint: Once again, the answer I'm looking for has something to do with *repeating* the experiment.] full credit

After repeating my experiment 100 times, we can expect that 95 of the confidence intervals generated will contain the true value.

"We are 95% confident that the 95th percentile confidence interval contains the true value of the parameter of interest."

c) (20 points each = 40 pts) Briefly describe two ways to construct confidence intervals for a parameter estimate? [I'm looking for one parametric and one non-parametric approach.]

1) Non parametric bootstrap (if not enough data or want to increase power, then parametric is also an option)
Sample data with replacement & estimate the parameter for each sample, repeat many times (k)
Using the distribution of parameter estimates, you can calculate the confidence interval using the quantile function or the standard error (2SE)

$$se_{boot} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\theta_i^* - \bar{\theta}^*\right)^2}$$

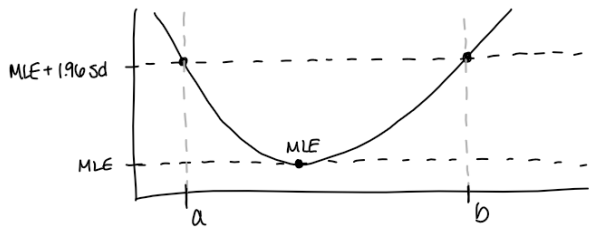
where θ_i^* is your sample parameter estimates ($\hat{\alpha}$ or $\hat{\beta}$)
& $\bar{\theta}^*$ is the average parameter estimates of the samples

$$\left(\bar{\theta}^* - se_{boot} z_{(1-\alpha/2)} \leq \theta \leq \bar{\theta} + se_{boot} z_{(1-\alpha/2)}\right) = 1-\alpha$$

2) Maximum likelihood estimate (parametric approach)
Start with the PDF, then find the likelihood function through the joint probability

Method 1
take the negative log, take a partial derivative of the NLL pertaining to α or β, & calculate the MLE
construct the confidence interval by deriving the analytical expression for the distribution, taking the desired quantiles of that distribution (say 2.5th & 97.5th for 2 tailed w/ 95% confidence) & adding/subtracting from the average parameter estimate to get the upper/lower bounds of the confidence interval

Method 2
calculate the x intercepts of the maximum likelihood function where the function is 1.96 (for 95% confidence) standard deviations more than the MLE



a = lower bound of CI
b = upper bound of CI