HW4_Brynestad_Cal

January 7, 2022

1 Homework 4: Confidence Intervals and Hypothesis Testing

Name: Cal Brynestad

This assignment is due on Canvas by **6:00PM on Friday November 5**. Your solutions to theoretical questions should be done in Markdown directly below the associated question. Your solutions to computational questions should include any specified Python code and results as well as written commentary on your conclusions. Remember that you are encouraged to discuss the problems with your classmates, but **you must write all code and solutions on your own**.

NOTES:

- Any relevant data sets should be available in the Homework 04 assignment write-up on Canvas. To make life easier on the grader if they need to run your code, do not change the relative path names here. Instead, move the files around on your computer.
- If you're not familiar with typesetting math directly into Markdown then by all means, do your work on paper first and then typeset it later. Remember that there is a reference guide linked on Canvas on writing math in Markdown. **All** of your written commentary, justifications and mathematical work should be in Markdown.
- Because you can technically evaluate notebook cells is a non-linear order, it's a good idea to do Kernel → Restart & Run All as a check before submitting your solutions. That way if we need to run your code you will know that it will work as expected.
- It is **bad form** to make your reader interpret numerical output from your code. If a question asks you to compute some value from the data you should show your code output **AND** write a summary of the results in Markdown directly below your code.
- This probably goes without saying, but... For any question that asks you to calculate something, you **must show all work and justify your answers to receive credit**. Sparse or nonexistent work will receive sparse or nonexistent credit.

---

[1]:
```
import pandas as pd
import numpy as np
from math import factorial
import matplotlib.pylab as plt
```

```
from scipy.special import binom
from scipy.stats import poisson

import scipy.stats as stats

%matplotlib inline
```

1.0.1 [20 points] Problem 1 - Sea Level

You have been contacted by the nation of Bangladesh, to assess whether there is statistical evidence for sea-level rise in the Cox's Bazar area.

You obtain from the University of Hawaii Sea Level Center's gigantic repository of sea-level data the daily mean sea levels file for the area. This data can be found in the `CoxBazaar.csv` file that is posted on Canvas. You may read more about the data here

In this problem, you will: 1. practice calculating confidence intervals, 1. practice wrangling a real-life data set into a form where you can actually compute these confidence intervals, because life will rarely be so kind as to simply hand you a nicely packaged and cleaned set of data, and 1. save Cox's Bazar in Bangladesh from a watery fate?

Part A: Read in the data file. Note the lack of a header. Create a header (using python and pandas) that labels the first column 'Year', the second column 'Month', the third column 'Day', and the last column 'Sea Level'.

[2]:
```
# Solution
file_path = 'CoxBazaar.csv'

df = pd.read_csv(file_path, names=['Year', 'Month', 'Day', 'Sea Level'])

df.head()
```

[2]:

	Year	Month	Day	Sea Level
0	1983	1	1	1638
1	1983	1	2	1592
2	1983	1	3	1475
3	1983	1	4	1417
4	1983	1	5	1445

Part B: Write a function `clean_data` to: 1. Take in a single argument of a raw sea level data frame (e.g., `dfCB` above), 2. Determine what the fill-value used to replace missing sea level (Sea Level)

data is, 3. Use the Pandas `DataFrame.dropna` method to remove all missing rows of data (use the fill-value to help do this), 4. select only the data point on the first day of each month, and 4. return a cleaned Pandas data frame.

Use your shiny new function to clean the `dfCB` data frame and save the results in a new data frame.

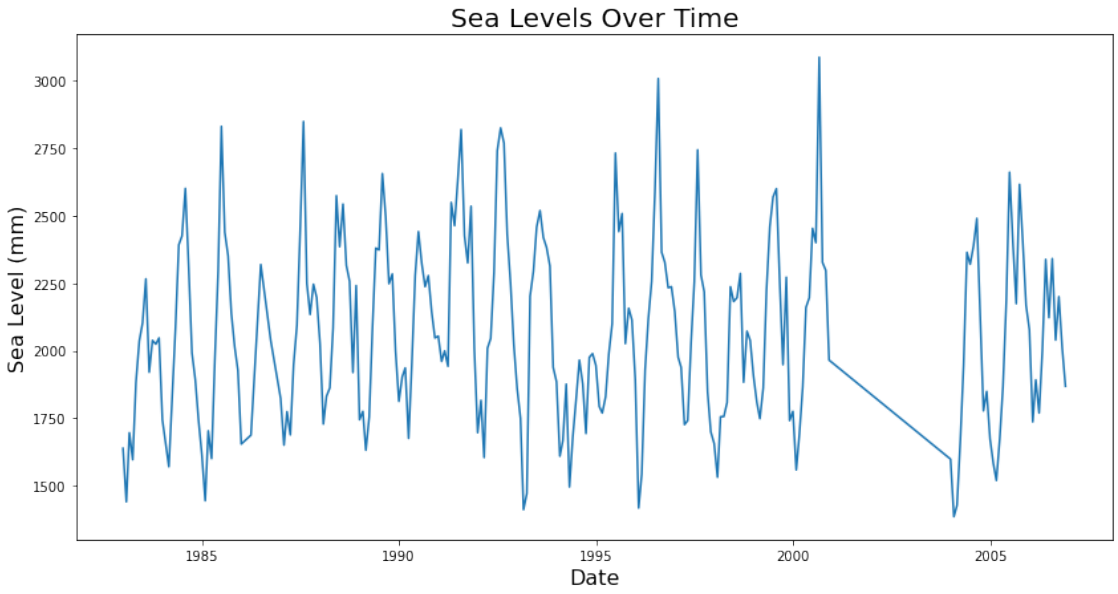
There is a very specific reason to sample only one daily data point per month. We will talk about it later.

[3]:
```
def clean_data(df):
    dfCB = df.replace(-32767, np.nan)

    dfClean = dfCB.dropna()

    dfFirsts = dfClean.loc[df['Day'] == 1]

    return dfFirsts




dfCleaner = clean_data(df)
```

Part C: Plot the cleaned time series of sea levels. Be sure to label your axes, including units. The UHSLC data includes the metadata accompanying our data set (linked in part A); if you are not sure about units, that would be a good place to start looking. For the x-axis, place the x tick marks on January 1 of each year that is divisible by 5 (i.e., 1980, 1985, ...), and label with that year. You may need to do additional processing in order to grab these indices.

[4]:
```
fig, ax = plt.subplots(figsize=(14,7))

labels = dfCleaner.index[(dfCleaner['Month'] == 1) & (dfCleaner['Year'] % 5 ==␣
 ↪0) & (dfCleaner['Day'] == 1)]
dfCleaner['Sea Level'].plot()
ax.set_xticks(labels)
ax.set_xticklabels(dfCleaner.loc[labels, "Year"])

ax.set_title("Sea Levels Over Time", fontsize=20)
ax.set_xlabel("Date", fontsize=16)
ax.set_ylabel("Sea Level (mm)", fontsize=16)
```

[4]:
```
Text(0, 0.5, 'Sea Level (mm)')
```



Part D: If you've plotted the data correctly, you'll notice a strange negatively sloped line between the years 2001 and 2005. Explain why this occurs.

Solution: A portion of data between 2001 and 2005 was eliminated from the data set because it didn't contain values for sea level.

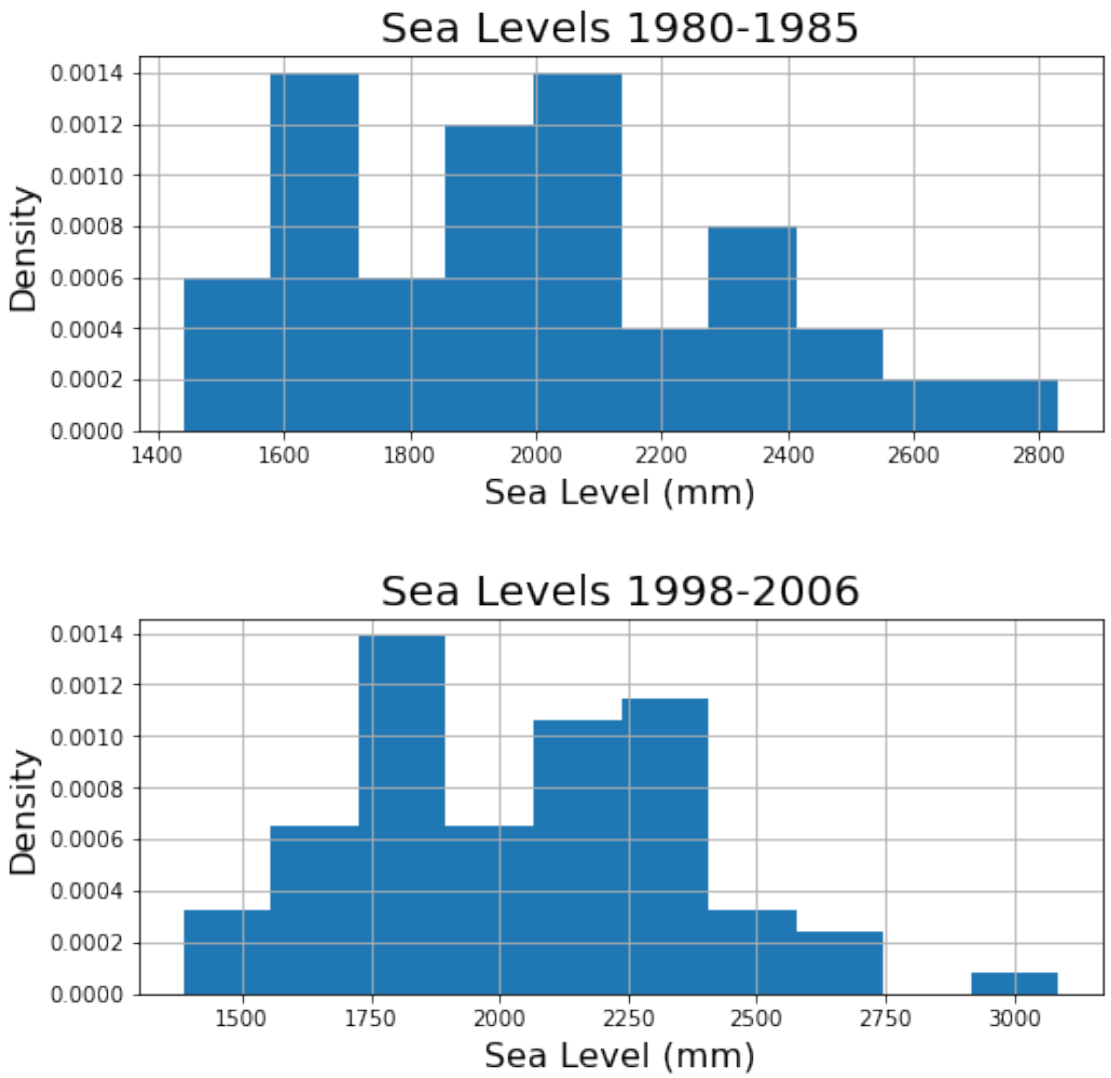
Part E: Use your cleaned sea levels data frame to create two new Pandas data frames or series: 1. one object to contain the sea levels between (and including) the years 1980 and 1985, and 2. another object to contain the sea levels between (and including) the years 1998 and 2006. Note, since there is no data for the years 2001-2003, each of these time periods covers 6 years.

Then, create a single-panel figure that includes density histograms of each period of sea levels. Be sure to label everything appropriately.

Finally, based on the data in front of you, formulate and state a hypothesis about how the mean sea level in the years 1980-1985 compares to the mean sea level in the years 1998-2006.

[5]:
```
# Solution:
dfBoxing1980_1985 = dfCleaner[(dfCleaner['Year'] >= 1980) & (dfCleaner['Year']␣
 ↪<= 1985)]
dfBoxing1998_2006 = dfCleaner[(dfCleaner['Year'] >= 1998) & (dfCleaner['Year']␣
 ↪<= 2006)]

fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(8,8))

dfBoxing1980_1985.hist(column='Sea Level', ax = axes[0], density = True)
dfBoxing1998_2006.hist(column='Sea Level', ax = axes[1], density = True)

axes[0].set_title("Sea Levels 1980-1985", fontsize=20)
```

```
axes[0].set_xlabel("Sea Level (mm)", fontsize=16)
axes[0].set_ylabel("Density", fontsize=16)

axes[1].set_title("Sea Levels 1998-2006", fontsize=20)
axes[1].set_xlabel("Sea Level (mm)", fontsize=16)
axes[1].set_ylabel("Density", fontsize=16)

fig.subplots_adjust(hspace=.5)
# axes[0].set_xlim([1400,3250])
# axes[1].set_xlim([1400,3250])
```



Based on the data I hypothesize that the mean sea level in the years 1980-1985 is roughly the same as the mean sea level in the years 1998-2006.

Part F: Compute a 95% confidence interval for each of (1) the mean sea level in the 1980-1985

time span ($\mu_{1980-1985}$) and (2) the mean sea level in the 1998-2006 time span ($\mu_{1998-2006}$). You may use Python for arithmetic operations and executing the calculations, but the relevant steps/set-up should be displayed in Markdown/MathJax.

Based on these two confidence intervals, do you think there is sufficient evidence to conclude that there is or is not a significant difference in the mean sea level between 1980-1985 and 1998-2006? Justify your answer.

Solution: For a 95% confidence interval we have that:

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

1.96 is the critical Z value for a confidence interval of 95%. We calculate our standard error using the sample size n and σ from the relevant data and use the same data to calculate $\bar{X}$. Plugging this information into the equation above using Python we get our confidence intervals for both time periods. From the confidence intervals calculated below I don't think that there's sufficient evidence to conclude that there is or is not a significant difference in the mean sea level between 1980-1985 and 1998-2006. This is because the confidence intervals share a common interval from 1966 to 2087 where a common sea level could exist but both confidence intervals also have possible true means that lie outside of their shared interval, so we can't tell if the two time periods true mean sea levels vary greatly or not.

[6]:
```
# Code here
n1=len(dfBoxing1980_1985)
xbar1=sum(dfBoxing1980_1985['Sea Level'])/n1
svar1=sum((xi-xbar1)**2/(n1-1) for xi in dfBoxing1980_1985['Sea Level'])
sd1=np.sqrt(svar1)
critz=stats.norm.ppf(.975)
print('The CI for the mean sea level in the 1980-1985 time span: ',␣
 ↪xbar1-critz*sd1/np.sqrt(n1),xbar1+critz*sd1/np.sqrt(n1))

n=len(dfBoxing1998_2006)
xbar=sum(dfBoxing1998_2006['Sea Level'])/n
svar=sum((xi-xbar)**2/(n-1) for xi in dfBoxing1998_2006['Sea Level'])
sd=np.sqrt(svar)
critz=stats.norm.ppf(.975)
print('The CI for the mean sea level in the 1998-2006 time span: ',␣
 ↪xbar-critz*sd/np.sqrt(n),xbar+critz*sd/np.sqrt(n))
```

```
The CI for the mean sea level in the 1980-1985 time span:  1870.1395446141344
2087.916010941421
The CI for the mean sea level in the 1998-2006 time span:  1966.9429086783366
2122.390424654997
```

Part G: Compute a 95% confidence interval for the ***difference in mean sea level*** between the 1980-1985 and the 1998-2006 time spans ($\mu_{1998-2006} - \mu_{1980-1985}$). Based on this, make a conclusion regarding your hypothesis from **Part E**, and compare to what your results in **Part F** implied. You may use Python for arithmetic operations and executing the calculations, but the relevant steps/set-up should be displayed in Markdown/MathJax.

Solution: For a 95% confidence interval for the difference in two means we have that:

$$(\bar{X} - \bar{Y}) \pm 1.96 \cdot \sqrt{\frac{\sigma_1^2}{\sqrt{m}} + \frac{\sigma_2^2}{\sqrt{n}}}$$

Where $\bar{X}$ is $\mu_{1998-2006}$ and $\bar{Y}$ is $\mu_{1980-1985}$ We have calculated each of these variables above so now we just plug into python:

[7]:
```
# Code Here
var1=svar1/n1
var2=svar/n
var_pool = var1 + var2
print("For a CI of: ", xbar1-xbar-critz*np.sqrt(var_pool),xbar1-xbar+critz*np.
 ↪sqrt(var_pool))
```

```
For a CI of:  -199.4209126339664 68.14313485618857
```

So we have that $-199.42 < \mu_{1998-2006} - \mu_{1980-1985} < 68.14$

We can say with 100% confidence that 95% of intervals generated in this manner show that the mean sea levels during the two time periods are no different on average, based on the sample data, as the CI contains 0. This provs our hypothesis from part E to be true. This is also what our result in part F implied could possibly be true as the two intervals computed there shared a common interval.

Part H: The confidence intervals from **Parts F** and **G** were derived using the Central Limit Theorem. Which assumption of the Central Limit Theorem would likely be violated if we took more than one measurement per month to form our samples, and why?

Solution: Assumptions of the CLT include that samples are independent of each other. Sampling each first day of each month of each year gives us an independent and identically distributed sample, as desired. If we didn't do so this would not be the case as the samples wouldn't be identically distributed and more importantly if 2 days were sampled that are right next to eachother, such as January 1, 1983 and January 2, 1983, these two sample points wouldn't be independent as knowing the sea level on January 1 gives us a pretty good idea of what the sea level on January 2 will be, and so these two data points aren't reaaly independent.

1.0.2 [20 points] Problem 2 - Hypothesis Testing

Pessimist Pete is waiting for the Buff Bus, and is rather impatient.

His friend tells him that busses arrive according to an exponential distribution with parameter $\lambda = 1/8$ (busses/min) for a *mean* waiting time of 8 minutes. Pete has been waiting for a while, and wants to prove their hypothesis wrong: he's thinking the wait is longer than that (one-tailed).\

attachment:buffbusbad.png

(Note: You may do calculations in Python if you wish, but all exposition should be in markdown.)

Part A: State the null and alternative hypotheses being tested.

Solution: H_0: $\mu = 8$, or, the mean waiting time for the bus is 8 minutes H_1: $\mu > 8$, or, the mean waiting time for the bus is greater than 8 minutes.

Part B: Devise a critical region test of the form "reject if $X > c$" where c is how long you have to wait until the bus arrives. Use a significance of $\alpha = .05$. How long do you wait before you reject the null hypothesis with a probability of type I error of 5%? **Round your answer to the nearest minute.**

Solution: We want to find the waiting time c, for how long we must wait for the bus in order to reject the null hypothesis at the .05 significance level. To do this we will solve: $P(X > c) = .05$ is equivalent to $1 - P(X \leq c) = .05$. We can evaulate this using the CDF of an exponential. The CDF of an exponential is $\int_0^x \lambda^{-\lambda t} dt = (-e^{-\lambda t})$ evaluated from 0 to x which $= 1 - e^{-\lambda t}$. So we have $1 - P(X \leq c) = .05$ is equivalent to $e^{-\lambda c} = .05 \rightarrow c = \frac{\ln .05}{-\lambda} \rightarrow c = \frac{\ln .05}{-\frac{1}{8}} \rightarrow c = -8 \ln .05$, which is equal to about 24 minutes. Therefore, we reject the null at the .05 significance level if we wait more than 24 minutes.

Part C: Suppose you observe a bus arrive 20 minutes after the previous bus. Again using a significance level of $\alpha = 0.05$ and your hypothesis from **Part A**, what would you conclude aboout the mean wait time?

Solution: Given that the bus arrives 20 minutes after the previous bus, and that we have shown above that we only reject the null hypothesis at the .05 significance level if we have to wait more than 24 minutes for a bus to arrive after the previous one left, and given that 20 < 24, in this case we fail to reject the null hypothesis and conclude that the mean wait time is $\mu = 8$.

Part D: Now perform a **p-value hypothesis test** using a significance level of $\alpha = 0.1$. Using the same hypotheses from **Part A**, and supposing that you wait 25 minutes for the next bus to arrive, what is your conclusion?

Solution: To find the p-value for the case where we wait 25 minutes we can use the CDF of an exponential distribution to help us find $P(X \geq 25)$. $P(X \geq 25) = 1 - P(X < 25) = 1 - (1 - e^{-\lambda \cdot 25}) = e^{\frac{-25}{8}}$ using our $\lambda = 1/8$. $e^{\frac{-25}{8}} = .0439$ This p-value represents the probability we wait 25 or more minutes given that the average wait time is 8 minutes. Since our calculated p-value is less than the significance level of .1, we reject the null hypothesis at the .1 significacne level in favor of the alternate hypothesis. Therefore, we have significant evidence that the wait time between busses is greater than 8 minutes.

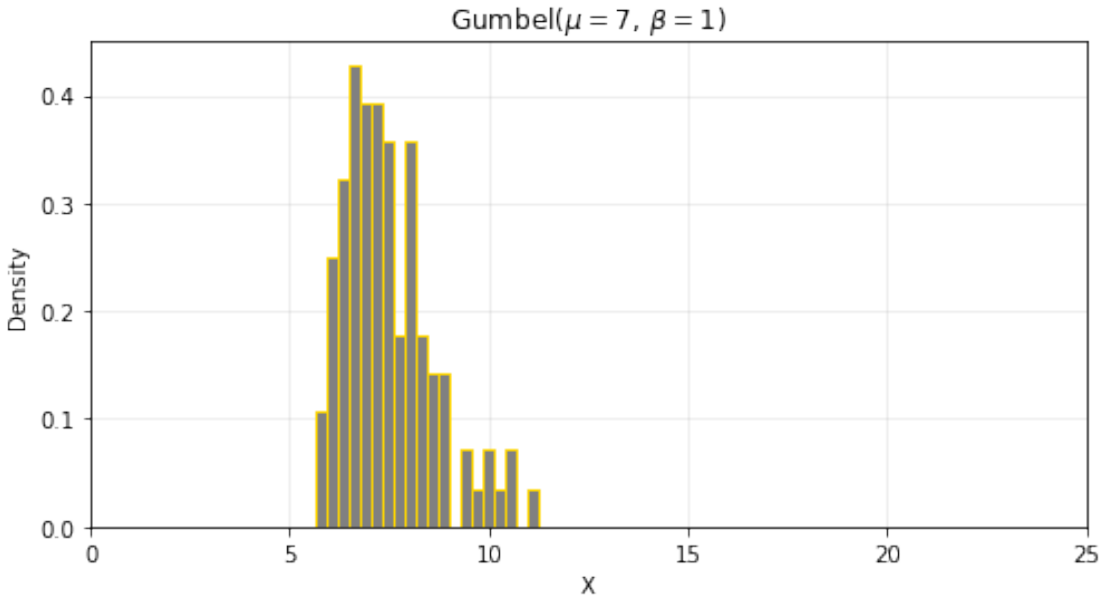
1.0.3 [20 points] Problem 3 - Confidence Intervals - Exploring the Theory

---

The Gumbel distribution is one of several distributions frequently used to model environmental extremes (for example, extreme temperatures and sea levels). It is also fairly asymmetric, and thus interesting for investigating confidence intervals. It is implemented in scipy.stats as gumbel_r, where the appendix "_r" denotes the right-skewed version of the Gumbel distribution (as opposed to the left-skewed).

Part A: Execute the following code cell to plot a histogram of 100 realizations from the Gumbel distribution with parameters $\mu = 7$ and $\beta = 1$. Be sure to leave this cell executed before turning in your assignment! Make your histogram grey with gold edges.

[8]:
```
mu = 7
beta = 1
n_sample = 100
x = stats.gumbel_r.rvs(loc=mu, scale=beta, size=n_sample)

fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(8,4))
plt.hist(x, color="grey",edgecolor="gold", bins=20, density=True)
ax.grid(alpha=0.25)
ax.set_axisbelow(True)
ax.set_xlabel('X')
ax.set_ylabel('Density')
ax.set_title(r'Gumbel($\mu=${}, $\beta=${})'.format(mu,beta))
plt.xlim([0,25])
plt.show()
```



Part B: Look up the analytical mean and variance of the Gumbel distribution with parameters $\mu = 7$ and $\beta = 1$ and calculate them here by hand. Note that the Euler–Mascheroni constant can be accessed via `np.euler_gamma`.

Use the empirical mean from your sample in **Part A**, and the true variance of the Gumbel distribution to compute by hand a 90% confidence interval for the mean.

Solution: $E(X) = \mu + (\gamma \cdot \beta) = 7 + \gamma = 7.58$ $Var(X) = (\frac{\beta \cdot \pi}{\sqrt{6}})^2 = \frac{\pi}{\sqrt{6}} = 1.65$ $Std(X) = \frac{\beta \cdot \pi}{\sqrt{6}} = 1.28$ Z critical for a 90% confidence interval is 1.64. Lower half of confidence interval: $7.58 - (1.64 *$

$(\frac{1.28}{\sqrt{100}})$) Upper half of confidence interval: $7.58 + (1.64 * (\frac{1.28}{\sqrt{100}}))$ Using exact values we get the 90% confidence interval for the mean from the code below:

[9]:
```
# Code Here
mean = 7 + (np.euler_gamma * 1)
sd = np.pi/np.sqrt(6)
var = sd**2

alph = .1
zcrit = stats.norm.ppf(1 - (alph/2))
L = mean - (zcrit * (sd/np.sqrt(100)))
H = mean + (zcrit * (sd/np.sqrt(100)))

print("The confidence interval is: ",L, "to", H)
print('The mean is: ', mean)
print('The variance is: ', var)
```

```
The confidence interval is:  7.3662549909127595 to 7.788176338890306
The mean is:  7.577215664901533
The variance is:  1.6449340668482264
```

Part C: A theoretical interlude. When one of the course CAs ran his solution code for **Part B**, he obtained a 90% confidence interval of $[7.340, 7.762]$ for the mean of the $Gum(\mu = 7, \beta = 1)$ distribution. For each of the following, explain why or why not the situation described is correct, given the technical definition of a 90% confidence interval we went over in class.

(i) If you had no other evidence regarding true mean of the $Gum(\mu = 7, \beta = 1)$ distribution, you could say there is a 90% chance that its true mean falls between 7.340 and 7.762.

(ii) If a class of 100 students all construct 90% confidence intervals for the mean of the $Gum(\mu = 7, \beta = 1)$ distribution, then we expect about 90 of their CIs to contain the true mean, and about 10 of them to miss the true mean.

(iii) There is a 90% probability that any given random variable sampled from $Gum(\mu = 7, \beta = 1)$ will be between 7.340 and 7.762.

Solutions: i) This is false, rather we are 90% confident that the true mean falls between 7.340 and 7.762. Or, we are 100% confident that 90% of CI's constructed in this manner will contain the true population mean, this being one of those CI's. ii) True, when we construct multiple 90% CIs, we are 100% confident that 90% of CI's constructed in this manner will contain the true population mean, given we construct enough of them. iii) False, this can be seen as false if you imagine having two different CIs. If we say that there is a 90% probability that any given random variable sampled from $Gum(\mu = 7, \beta = 1)$ will be between the first CI, and the second CI was contructed in the same way but has different values, if what we said was true for the first CI we'd be able to say the same for the second CI, that there is a 90% probability that any given random variable sampled from $Gum(\mu = 7, \beta = 1)$ will be between the second CI as well. However clearly this makes no sense as they are two different CIs so the probability that any given random variable sampled from $Gum(\mu = 7, \beta = 1)$ is between CI 1 and CI 2 can't be the same. Again, rather we are 100% confident that 90% of CI's constructed in this manner will contain the true population mean.

Part D: In this part you'll write a function to investigate the *coverage properties* of a confidence

interval for the mean of the Gumbel distribution. Complete the following function to randomly sample $m = 500$ sample means with sample size $n = 100$ for the Gumbel distribution with parameters $\mu = 7$ and $\beta = 1$. For each random sample, compute the 80% confidence interval for the mean. Note that you actually know that the variance for the true population distribution is, σ^2. Your function should do two things:

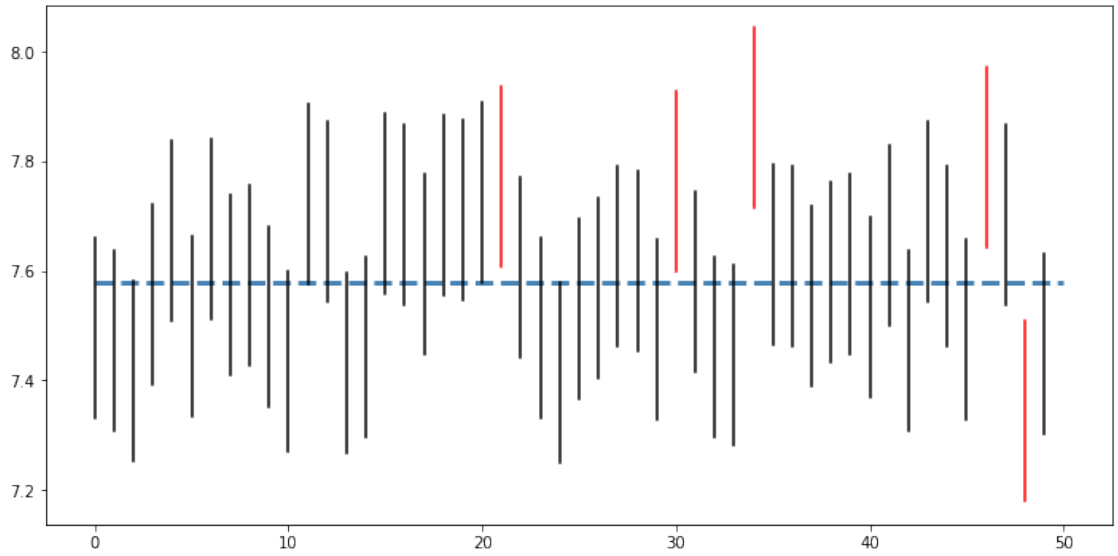
1. Report the proportion of confidence intervals that successfully cover the true mean of the distribution
2. Make a plot of 50 randomly selected confidence intervals. Overlay the intervals on the line $y =$ True mean (from **Part B**). Color confidence intervals black if they cover the true mean, and red if they don't.

Be sure to leave this cell executed before turning in your assignment!

[10]:
```
def confidence_intervals(m=500, n=100):
    mean,var = stats.gumbel_r.stats(loc=7, moments='mv')

    alpha = .2
    count = 0
    CIs = []

    for i in range(m):
        x = stats.gumbel_r.rvs(loc=7,size=n)
        xbar = np.mean(x)
        z_alpha = stats.norm.ppf(1-(alpha/2))
        SE = np.sqrt(var)/np.sqrt(n)
        CImin = xbar-(z_alpha*SE)
        CImax = xbar+(z_alpha*SE)
        CII = (CImin, CImax)
        CIs.append(CII)

        if (CImin<mean and CImax>mean):
            count+=1

    fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(12,6))
    ax.plot([0,50], [mean, mean], color="steelblue", ls='--', lw=3)

    for x in range(50):
        if CIs[x][0] < mean and CIs[x][1] > mean:
            ax.plot([x,x], [CIs[x][0], CIs[x][1]], color='black')
        else:
            ax.plot([x,x], [CIs[x][0], CIs[x][1]], color='red')


    return count/m
```

```
answer = confidence_intervals()
print('The proportion of confidence intervals that successfully cover the true␣
 ↪mean of the distribution is:', answer)
```

```
The proportion of confidence intervals that successfully cover the true mean of
the distribution is: 0.85
```



Part E: Does the proportion of confidence intervals that cover the true mean of the distribution agree with the theory described in class? Justify your conclusion.

Solution: Yes, by using an 80% confidence interval, we are 100% confident that 80% of the CIs constructed will contain the value of the true mean, which is the case.

1.0.4 [20 points] Problem 4 - Confidence Intervals with Proportions

Part A: You work for an engineering firm that has been hired to construct a bridge from the Engineering Center to the Physics building. Thousands of students will walk along this bridge each day, so structural failure means injury or death for many people.

You are in charge of quality control for the average strength of carbon fiber that will be used to construct the bridge. Thinking back fondly to your days in CSCI 3022, you set up a hypothesis test in which your alternative hypothesis is that the strength of the carbon fiber is below tolerance, and therefore unsafe. What is the null hypothesis? Would you rather have a low Type I error rate or a low Type II error rate? Explain.

Solution: The null hypothesis H_0 is that the strength of the carbon fiber is of the proper tolerance and therefore safe. We would rather have a low Type II error rate. A type II error would mean that H_0 is false, or that the strength of the carbon fiber is not of the proper tolerance, and that we fail to reject H_0. In other words the bridge is unsafe, and we fail to recognize that the bridge is unsafe. This is worse than a type I error as a high type II error rate would mean that the bridge is

potentially unsafe where as a type I error would mean than the bridge is safe but we think it isn't, which would only lead to more precaution that's not necessarry.

Part B: Gandalf, the famous wizard data scientist, is working for the same engineering firm as you. He is a legend around the office! Word around the water cooler is that out of all of the 95% confidence intervals that Gandalf has constructed, 870 of them have turned out to actually capture the true population mean. Since Gandalf is a data science wizard and you can be sure he is constructing his confidence intervals correctly and collecting and using his data honestly, about how many 95% confidence intervals would you expect him to have constructed total? Explain your reasoning fully with words as well as some math.

Solution: The construction of 95% confidence intervals means that we are 100% confident that 95% of the intervals constructed contain the true population mean. Given, Gandalf is perfect in his constructions this means that 870 confidence intervals is 95% of the total amount of confidence intervals he has constructed. So to get the total number of confidence intervals that he has constructed: $\frac{95}{100} = \frac{870}{x} \rightarrow 95x = 87000 \rightarrow x = 915.7894737$ Therefore, the total number of confidence intervals that Gandalf has constructed is (rounding up so he's not so perfect after all) 916.

Part C: In general, which is wider: a 95% confidence interval or a 99% confidence interval?

Why? Provide a brief explanation.

Solution: A 99% confidence interval is wider. Constructing a 99% CI in comparison to a 95% CI means that we are more confident that the true population parameter is in our 99% CI. So it makes sense that the 99% CI would be wider as it would include more values inside of it and therefore it is more likely that a calculated sample statistic lies inside of the interval, making us more confident.

Part D: Gandalf decides to run a study on magic wands. He observes a sample of 73 black wands and find that 49 of them are in good working order. Then, he observes a sample of 58 brown wands and find that 51 of them are in good working order.

Is there statistical evidence at the 0.05 significance level that the true proportion of brown wands that are in good working order is 0.1 higher than the true proportion of black wands that are in good working order? Perform a test that computes and properly interprets a p-value.

Solution: H_0 : The true proportion of brown wands that are in good working order is exactly 0.1 higher than the true proportion of black wands that are in good working order. H_1 : The true proportion of brown wands that are in good working order is greater than 0.1 higher than the true proportion of black wands that are in good working order.

[20]:
```
# Code Here
#alpha = .05
#zcrit = stats.norm.ppf(1 - .05) #one tailed test
#L = (p2 - p1) - (zcrit*se)
#H = (p2 - p1) + (zcrit*se)

#print('The 95% confidence interval is:',L,H)


m = 73
```

```
n = 58

p1 = 49/73
p2 = 51/58

se = np.sqrt((p1*(1-p1)/m) + (p2*(1-p2)/n))

p3 = p2-p1

ans = (p3 - .1)/se

pval = 1 - stats.norm.cdf(ans)

print("P Value is:", pval)
```

```
P Value is: 0.060394950555197724
```

The p value obtained for the likelihood that the difference in proportions between the black and brown wands being $\frac{51}{58} - \frac{49}{73} = .208$, given that the real difference is .1, is .06. This means that if the true proportion of brown wands that are in good working order is 0.1 higher than the true proportion of black wands that are in good working order, the likelihood of us obtaining the results we did from our samples of wands is about 6%. Since we are testing at the .05 significance level and .06 is greater than .05, there IS NOT statistical evidence for H_1 at the 0.05 significance level and therfore we conclude that the true proportion of brown wands that are in good working order is 0.1 higher than the true proportion of black wands that are in good working order

1.0.5 [20 Points] Problem 5 - Hypothesis Testing - Theory

You are working as a Data Scientist. You decide to let one of your coworkers, Bob, do some hypothesis testing for you. Unfortunately, Bob is not a master of logic and inference and many mistakes are made throughout the day as the two of you team up to tackle some inference work. In each case, clearly explain why the hypothesis testing setup or conclusion is incorrect.

Part A: There is some data on the characteristics of customers that visited your company's website over the previous month. Bob wants to perform an analysis on the proportion of last month's website visitors that bought a product. Let X be the random variable describing the number of website visitors who bought a product in the previous month, and suppose that the population proportion of visitors who bought a product is p. Bob is interested to see if the data suggests that more than 20% of website visitors actually buy a product. He decides to perform the test with a null hypothesis of $H_0 : \hat{p} = 0.20$.

Solution: If X is the number of website visitors who bought a product in the previous month, then $\hat{p} = \frac{X}{n}$, where n is total website visitors for that month. So $\hat{p}$ is the test statistic. Therefore the null hypothesis should be $H_0 : p = 0.20$, not $H_0 : \hat{p} = 0.20$, as the null hypothesis should hypothesize about the true proportion of website visitors who buy a product, not about the test statistic.

Part B: Bob decides instead to do his hypothesis test with a null hypothesis of $H_0 : p < 0.20$.

Solution: The null hypothesis is usually of the form: H_0: $\theta = \theta_0$. In this case it should be $H_0 : p = 0.20$. This is because Bob is interested to see if the data suggests that more than 20% of website

visitors actually buy a product, so it is irrelevant whether the true population proportion is .2 or .19 or .18. We want to assume that the null is true, $H_0 : p = 0.20$, and use this to calculate a test statistic from a sample potentially showing that the true population proportion is higher than .2, or to reject the null in favor of $H_1 : p > 0.20$.

Part C: Now Bob is finally on track with reasonable hypotheses of $H_0 : p = 0.20$ and $H_1 : p > 0.20$. He computes for the sample proportion a normalized test-statistic of $z = 2.4$ and states that since $z = 2.4 > 0.01$ there is insufficient statistical evidence at the $\alpha = 0.01$ significance level to conclude that the proportion of customers who buy a product is greater than 20%.

Solution: Bob forgot to compute a Z critical value based on his chosen $\alpha = 0.01$. To compute the Z critical value at the $\alpha = 0.01$ significance level, Bob must find the corresponding Z scores that lie in the rejection region for the chosen significance level. This calculation is done below to obtain a Z critical of 2.33. This means that if Bob obtains a Z score greater than 2.33 using his sample proportion, then Bob can reject the null hypothesis at the .01 significance level. Since Bob's calculated normalized test-statistic of z is 2.4 which is greater than 2.33, Bob can in fact reject the null hypothesis at the .01 significance level.

[13]:
```
stats.norm.ppf(.99)
```

[13]:
```
2.3263478740408408
```

Part D: Bob is again conducting the hypothesis test of $H_0 : p = 0.20$ and $H_1 : p > 0.20$, and found the test-statistic $z = 2.4$. He computes his p-value as:

$$\text{p-value} = \texttt{stats.norm.cdf(2.4)} = 0.9918$$

Since his p-value (0.9918) is greater than the significance level (0.01), he again claims that there is insufficient evidence at the 1% significance level to conclude that the proportion of customers who buy a product is greater than 20%.

Solution: Bob calculated his p-value incorrectly. The correct p-value is $p(Z \geq 2.4) = 1 - p(Z \leq 2.4) = 1 - .992 = .008$. This is the correct p-value representing the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming the null hypothesis is correct. What Bob calculated was the the probability of obtaining results not as extreme. With the correct p-value of .008, the correct conclusion is that since the calculated p-value of .008 is less than the significance level (0.01), there is sufficient evidence at the 1% significance level to conclude that the proportion of customers who buy a product is greater than 20%.

[14]:
```
stats.norm.cdf(2.4)
```

[14]:
```
0.9918024640754038
```

Part E: Bob is again conducting the hypothesis test of $H_0 : p = 0.20$ and $H_1 : p > 0.20$. Suppose he computes a p-value of 0.03, and then concludes that there is only a 3% probability that the null hypothesis is true.

Solution: The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. So what the p-value of .03 actually tells Bob is that assuming the population proportion of .2 is in fact true,

there is only a 3% probability of him obtaining the results he did in his statistical hypothesis test. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

[ ]: