Hw7

gjd9961

November 2021

1. (Short questions) Justify all your answers mathematically.

   (a) For any random variable $\tilde{a}$, can $E^2(\tilde{a})$ be smaller than $E(\tilde{a}^2)$?

   Yes, this can be true. We know that

   $$E(\tilde{a}^2) - E^2(\tilde{a}) = E((\tilde{a} - E(\tilde{a}))^2)$$

   Which can be expressed as

   $$E((\tilde{a} - E(\tilde{a}))^2) = \sum_{a \in \mathbf{R}_{\tilde{a}}} p_{(\tilde{a}(a)) \times (a - E(\tilde{a}))^2}$$

   And since probabilities are non negative by definition, and the other term is being squared, we have two positive quantities multiplying each other and then getting summed. Therefore, we cannot have a negative value for $E((\tilde{a} - E(\tilde{a}))^2)$ which means that $E((\tilde{a} - E(\tilde{a}))^2) \geq 0$ and $E(\tilde{a}^2) - E^2(\tilde{a}) \geq 0$ and that $E^2(\tilde{a}) \leq E(\tilde{a}^2)$

   (b) If $\tilde{a}$ and $\tilde{b}$ have the same distribution and are independent, is it true that $E(\tilde{a}\tilde{b}) = E^2(\tilde{a})$?

   Yes this is true, and we can leverage the fact that they are independent and that if they have the same distribution they have the same expectation ($E(\tilde{a}) = E(\tilde{b})$) to show that $E(\tilde{a}\tilde{b}) = E^2(\tilde{a})$:

   $$E(\tilde{a} \times \tilde{b}) = E(\tilde{a}) \times E(\tilde{b}) = E(\tilde{a}) \times E(\tilde{a}) = E^2(\tilde{a}) \quad \square$$

   (c) A teacher of a class of n children asks their parents to leave a present under the Christmas tree in the classroom. The day after, each child picks a present at random. What is the expected number of children that end up getting the present bought by their own parents? (Hint: Define a random variable I_i that is equal to one when kid i gets the present bought by their own parents, and to zero otherwise.)

   Since there are n kids in the classroom, there will be n presents under the tree. Picking at random, any child will have a $\frac{1}{n}$ chance of getting the present their parents left. Well define the present a child pics as a random variable $\tilde{i}$ and when $\tilde{i} = 0$ then the child did not get the present their parent left, and when $\tilde{i} = 1$ then they did.

   $$P_{\tilde{i}}(\tilde{i} = 0) = 1 - \frac{1}{n} \text{ and } P_{\tilde{i}}(\tilde{i} = 1) = \frac{1}{n}$$

Since there are n kids and $P_{\tilde{i}}(\tilde{i}=1)=\frac{1}{n}$ then the expectation is calculated as follows:

$$E(\text{Kids that get their parents present}) = E(\sum_{i=1}^{n} E(\tilde{i})) = n \times \frac{1}{n} = 1$$

That means that on average, one child in the classroom should get the present their parents left under the tree by random chance alone.

2. (Computer) We model the time (in years) until a computer breaks down as a random variable $\tilde{t}$. The time depends on whether the computer has a defect or not, which is modeled by a random variable $\tilde{d}$. If the computer has a defect ($\tilde{d}=1$), then $\tilde{t}$ is an exponential with parameter 2. If it does not ($\tilde{d}=0$), then $\tilde{t}$ is an exponential with parameter 1. The probability that the computer has a defect is 0.1.

   (a) Is the conditional expectation of $\tilde{t}$ given $\tilde{d}$ a discrete or continuous random variable? What is its pmf or pdf?

   Since we're plugging a random variable into a continuos function, our conditional expectation of $\tilde{t}$ given $\tilde{d}$ is a continuos random variable. Its pdf can be described in the following way:

$$PDF = f_{t|d}(t|d) = \begin{cases} Exp(2) & \text{if } \tilde{d}=1 \\ Exp(1) & \text{if } \tilde{d}=0 \end{cases}$$

   (b) What is the variance of $\tilde{t}$?

   If we want to compute the variance of $\tilde{t}$ we must first compute the marginal

$$Marignal = \sum_{0}^{1} P_d df_{t|d}(t|d) = p_d(0) f_{t|0}(t|0) + p_d(1) f_{t|1}(t|1) = .9 \times e^{-t} + .1 \times e^{-2t}$$

   Now we can continue to compute the expectation for $\tilde{t}$ and $\tilde{t^2}$ and then the variance

$$E(t) = \int_0^\infty \tilde{t} f_t(t) = 0.95 \text{ and } E(t^2) = \int_0^\infty \tilde{t^2} f_t(t) = 1.85$$

   Finally:

$$Var(\tilde{t}) = E(t^2) - E(t)^2 = 1.85 - .95^2 = .0.9475$$

   (c) A company buys 100 of these computers. If the time until they break down is distributed as explained above, and they are all independent, what is the mean and variance of the number of computers that break down during the first year?

   We could model the nubmer of computers that breaks down during the first year by modeling the 100 computers in a binomial distribution. Our size would be 100, and our probability would be .65 as

$$f_{t|t,d}(t|t,d) = \int_0^1 f_t(t) = .65$$

   Thus the mean of the distribution would be $n \times p = 65$ and the **Variance** would be $n \times p(1-p) = 100 \times .65 \times (1-.65) = 22.75$

3. (Law of conditional variance) In this problem we define the conditional variance in a similar way to the conditional expectation.

   (a) What is the object $Var(\tilde{b}|\tilde{a} = a)$ (i.e. is it a number, a random variable or a function)? What does it represent?

   At the end of the day, it just represents a number, the variance of a random variable $\tilde{b}$ given some information about the random variable $\tilde{a}$. This is because once we observe a value of $\tilde{a}$, a becomes fixed, and then when we look at joint density of the two random variables and then condition on the event a we observed, we get a new distribution. Then the variance of the distribution is a fixed number, the mathematical object in question.

   (b) Setting $h(a) = Var(\tilde{b}|\tilde{a} = a)$ we define the conditional variance as $Var(\tilde{b}|\tilde{a}) = h(\tilde{a})$. What is this object?

   Now what we have is not a number, because $\tilde{a}$ is not observed and fixed to any given number, that is to say, we're not sure exactly what to condition on. We know that if we define a function in terms of random variables, the resulting function is a random variable, thus, the object we have on our hands is a random variable.

   (c) Prove the law of conditional variance:

$$Var(\tilde{b}) = E(Var(\tilde{b}|\tilde{a})) + Var(E(\tilde{b}|\tilde{a})) \tag{1}$$

   and describe it in words.

   We can prove this in the following way:

$$\begin{aligned} E(Var(\tilde{b}|\tilde{a})) &= E(E(\tilde{b}^2|\tilde{a})) - E(E(\tilde{b}|\tilde{a})^2)) \\ E(Var(\tilde{b}|\tilde{a})) &= E(\tilde{b}^2) - E(E(\tilde{b}|\tilde{a})^2) \\ &\text{and} \\ Var(E(\tilde{b}|\tilde{a})) &= E(E(\tilde{b}|\tilde{a})^2) - E(E(\tilde{b}|ra))^2 \\ Var(E(\tilde{b}|\tilde{a})) &= E(E(\tilde{b}|\tilde{a})^2) - E(\tilde{b})^2 \end{aligned} \tag{2}$$

   Using what we computed, we can substitute these identities in and what we now have is:

$$\begin{aligned} E(Var(\tilde{b}|\tilde{a}) + Var(E(\tilde{b}|\tilde{a})) &= E(\tilde{b}^2) - E(\tilde{b})^2 \\ E(\tilde{b}^2) - E(\tilde{b})^2 &= Var(\tilde{b}) \end{aligned} \tag{3}$$

   What we have shown is that the Expectation of the variance of a random variable given some information about another random variable plus the variance of the expectation of a random variable given some information about a random variable is equal to the variance of the primary random variable (not the one being conditioned on).

4. (Water salinity and temperature) In this question, we use oceanographic data to study the relationship between salinity and temperature in sea water. We perform our analysis on a cleaned and subsampled version of the data, `bottle.csv`. The script is available at https://github.com/cfgranda/prob_stats_for_data_science/blob/main/hw7/conditional_expectation_EXERCISE.ipynb

(a) Plot an estimate of the conditional mean of salinity given the temperature along with the scatter plot of data. Justify any choices you make.

(b) Annotate your plot to incorporate the conditional standard deviation of salinity given the temperature (apart from plotting the conditional mean, plot the conditional mean $\pm$ the conditional standard deviation).

(c) Do you expect your estimates to be equally reliable at every point? Please explain your reasoning. (We are not looking for a mathematical answer, you can just explain intuitively.)

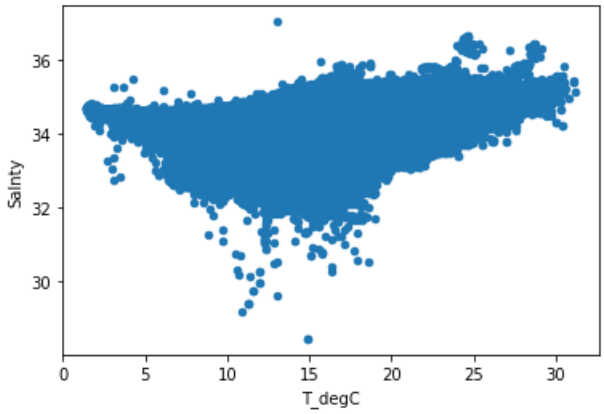
No, they definitely wont be reliable at every point. There were choices made about bucketing the data, and some buckets (such as from $0 < t < 6$ or even $25 < t < 30$) that might have less values than the others and therefore their means will be less reliable compared to other buckets containing more values (such as $10 < t < 20$).

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import random
import math
```

In [4]:
```
bottle = pd.read_csv('bottles.csv')
```

In [7]:
```
bottle.plot.scatter(x='T_degC', y='Salnty')
```

Out[7]: <AxesSubplot:xlabel='T_degC', ylabel='Salnty'>

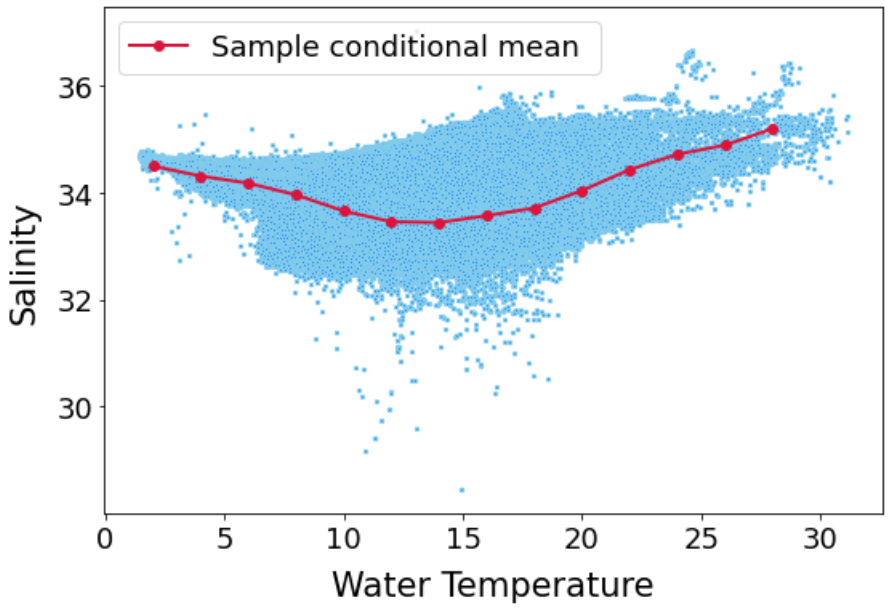


In [72]:
```
salnty = np.array(bottle['Salnty'])
temp = np.array(bottle['T_degC'])

salnty_clean =salnty[~np.isnan(temp)]
temp_clean = temp[~np.isnan(temp)]

temp_clean = temp_clean[~np.isnan(salnty_clean)]
salnty_clean = salnty_clean[~np.isnan(salnty_clean)]
```

In [113]:
```
max_val = np.max(temp[~np.isnan(temp)])
width_bin = 2
fig = plt.figure(figsize = (9,6))
plt.scatter(temp,salnty, s=5, c="dodgerblue", marker='o', edgecolor="skyblue")

# TODO: create bins from 0 to the maximum to discretize continous temperatures
grid = list(range(0,math.ceil(max_val),2))

# TODO: Compute the conditional expection of sanlinity given temperture
cond_average_salnty = np.zeros(len(grid))
cond_average_salnty_ind = np.zeros(len(grid))
cond_average_salnty_lists = dict()

i=0
while i < len(salnty_clean):

    index= math.floor(temp_clean[i]/2)
    cond_average_salnty[index]+=salnty_clean[i]
    cond_average_salnty_ind[index]+=1

    if index in cond_average_salnty_lists.keys():
        cond_average_salnty_lists[index].append(temp_clean[i])

    else:
        cond_average_salnty_lists[index]= [temp_clean[i]]

    i+=1

i=0
```

```
while i<len (cond_average_salnty_ind):
    cond_average_salnty[i] = cond_average_salnty[i]/cond_average_salnty_ind[i]
    i+=1



plt.plot(grid[1:-1],cond_average_salnty[1:-1],'-o',lw=2,color='crimson', label="Sample conditio
plt.ylabel("Salinity", fontsize=21,labelpad=10)
plt.xlabel("Water Temperature", fontsize=21,labelpad=10)

plt.legend(fontsize=18)
# plt.xlim(-5,30)
# plt.ylim(-12,23)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.gcf().subplots_adjust(bottom=0.15)
plt.gcf().subplots_adjust(left=0.15)
plt.savefig('conditional_expectation.pdf')
```

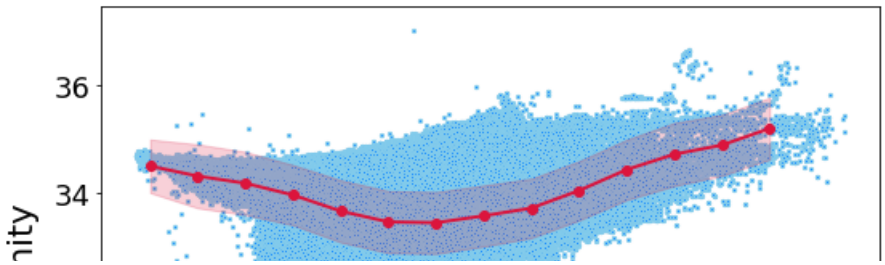
```
<ipython-input-113-9b9c79b847ec>:47: UserWarning: Creating legend with loc="best" can be slo
w with large amounts of data.
  plt.savefig('conditional_expectation.pdf')
<ipython-input-113-9b9c79b847ec>:47: UserWarning: Creating legend with loc="best" can be slo
w with large amounts of data.
  plt.savefig('conditional_expectation.pdf')
C:\Users\jonah\anaconda3\lib\site-packages\IPython\core\pylabtools.py:132: UserWarning: Crea
ting legend with loc="best" can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```



In [114]:
```
# TODO: Compute the conditional standard deviation of sanlinity given temperture
cond_std_salnty = np.zeros(len(grid))

i=0
while i<len (cond_average_salnty_ind):
    cond_average_salnty_lists[i] = np.std(cond_average_salnty_lists[i] )
    cond_std_salnty[i]=cond_average_salnty_lists[i]
    i+=1


fig = plt.figure(figsize = (9,6))
plt.scatter(temp,salnty, s=5, c="dodgerblue", marker='o', edgecolor="skyblue")


plt.plot(grid[1:-1],cond_average_salnty[1:-1],'-o',lw=2,color='crimson', label="Sample conditio
plt.fill_between(grid[1:-1], cond_average_salnty[1:-1]-cond_std_salnty[1:-1],
                 cond_average_salnty[1:-1]+cond_std_salnty[1:-1], color='crimson', alpha=0.2, la

plt.ylabel("Salinity", fontsize=21,labelpad=10)
```

```
plt.xlabel("Water Temperature", fontsize=21,labelpad=10)

plt.legend(fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.gcf().subplots_adjust(bottom=0.15)
plt.gcf().subplots_adjust(left=0.15)
plt.savefig('conditional_expectation_w_std.pdf')
```

```
<ipython-input-114-266567546691>:27: UserWarning: Creating legend with loc="best" can be slo
w with large amounts of data.
  plt.savefig('conditional_expectation_w_std.pdf')
<ipython-input-114-266567546691>:27: UserWarning: Creating legend with loc="best" can be slo
w with large amounts of data.
  plt.savefig('conditional_expectation_w_std.pdf')
C:\Users\jonah\anaconda3\lib\site-packages\IPython\core\pylabtools.py:132: UserWarning: Crea
ting legend with loc="best" can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```



In [107]:
```
cond_average_salnty = np.zeros(len(grid))
cond_average_salnty_ind = np.zeros(len(grid))
cond_average_salnty_lists = dict()

i=0
while i < len(salnty_clean):

    index= math.floor(temp_clean[i]/2)
    cond_average_salnty[index]+=salnty_clean[i]
    cond_average_salnty_ind[index]+=1

    if index in cond_average_salnty_lists.keys():
        cond_average_salnty_lists[index].append(temp_clean[i])

    else:
        cond_average_salnty_lists[index]= [temp_clean[i]]

    i+=1
```

In [104]:
```
i=0
while i<len (cond_average_salnty_ind):
    cond_average_salnty_lists[i] = np.std(cond_average_salnty_lists[i] )
    i+=1

i=0
while i<len(cond_std_salnty):

    cond_std_salnty[i]=cond_average_salnty_lists[i]
    i+=1
```

In [ ]: