Midterm CSCI ~~356~~ 443 Answers

Spring 2024

Useful equations

Some equations I wrote on the board at the beginning of the exam:

The mean of a set of samples drawn from a random variable X is given by

$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

The sample variation of samples drawn from X is given by

$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \left[(\sum_{i=1}^{n} x_i^2) - n\overline{x}^2 \right]$

The sample standard deviation of samples drawn from X is given by

$s_x = \sqrt{s_x^2}$

The Pearson correlation coeffidient is given by

$r_{xy} = \frac{s_{xy}}{s_x s_y}$

Types of Data

Data comes in several types including

categorical: the data can be divided into categories like species, color, garment size {S, M, L}

numerical: numeric data is comprised of numbers: counts, quantites, measurements.

Categorical data can be further divided into ordinal and nominal data.

ordinal data has a natural ordering. For example Small, Medium, Large sizes can be ordered from smallest to largest.

nominal data lacks a natural ordering. For example, "Truck" and "Car" are examples of nominal categorical data. We could order the words "Car" and "Truck" in alphabetical order, but ordering based on spelling of the labels themselves is usually considered insufficient to consider a type of data as ordinal since anything that has an assigned label in a written language that uses an alphabet can be ordered into an alphabetical ordering.

Problem 1 (10 points total, 1 point each)

Which of the following data types is numeric, and which is categorical? For the categorical data, specify if it is nominal or ordinal.

- a) Temperature in degrees Celsius.

  Answer: Numeric

- b) Rankings of a movie (e.g., Excellent, Good, Fair, Poor).

  Answer: Categorical (Ordinal)

- c) The colors of cars in a parking lot.

  Answer: Categorical (Nominal)

- d) The number of shares a company has sold.

  Answer: Numeric

- e) The pH level of soil samples.

  Answer: Numeric

- f) The zip codes of addresses in a city.

  Answer: Categorical (Nominal). Several students missed this question. A good rule of thumb: "No math? Not numeric." One does not usually perform mathematics on zip codes.

- g) The order of finishers in a race (1st, 2nd, 3rd, etc.).

  Answer: Categorical (Ordinal). As with 1(f), some stated this is numeric. Same rule of thumb.

- h) The types of payment methods used in transactions (e.g., cash, credit card, check).

  Answer: Categorical (Nominal)

- i) Age of participants in a study.

  Answer: Numeric, although it could be categorical (nominal) depending on how the age is represented. I accepted either answer.

- j) The breed of dogs in a kennel.

  Answer: Categorical (Nominal)

True / False

Problem 2 (10 points)

Circle whether a statement is true or false.

- a) True/False: A random experiment is an action or process that leads to one of many possible outcomes.

  Answer: True.

- b) True/False: A trial is the set of all possible outcomes in a random experiment.

  Answer: False. A trial is an individual performance of a random experiment. It isn't the set of all possible outcomes.

- c) True/False: I draw a card from a 52-card deck. The card is a 3 of spades. The 3 of spades is an outcome for a trial in which a single card is drawn.

  Answer: True.

- d) True/False: The number of pages in the first edition of "To Kill a Mocking Bird" by Harper Lee is a random variable.

  Answer: False. The number of pages in the first edition of "To Kill A Mockingbird" is a fixed number. It isn't a random variable.

- e) True/False: Drawing a single card from a standard 52-card deck constitutes a trial of a random experiment. Each card can be considered an individual outcome. Together, these 52 outcomes comprise the sample space for the random variable X, where X represents the event of drawing any one of the 52 cards.

  Answer: True. The sample space must contain every possible outcome. When drawing from a standard 52-card deck, the 52 cards constitute all possible outcomes. The drawing of a card is a single execution of a random experiment of drawing a card from a standard 52-card deck and is thus a trial.

- f) True/False: A trial is the actual performing of a random experiment.

  Answer: True.

- g) True/False: The outcome of a random experiment is always numerical.

  Answer: False. A outcome of a random experiment can be numerical, but it need not be. It could be categorical, or it could be

something that doesn't neatly fit into either numerical or categorical. For example it could be a mixed type such as customer data comprising both numerical (e.g., salary) and categorical data (e.g., gender).

- h) True/False: In a random experiment, the trial and the outcome are the same thing.

  Answer: False. The trial is executing the random experiment. The outcome is what happens.

- i) True/False: The bus arrives early on Monday, Tuesday, and Wednesday is an event.

  Answer: True. If you can assign a probability to it then it is an event.

- j) True/False: Counting the number of minutes in an hour is a random experiment.

  Answer: False. There are always 60 minutes in an hour; therefore, counting the numnber of minutes is not a *random* experiment.

The Dreemes Job

Dreemes is a seaside town known for its beach. The Dreemes Chamber of Commerce has hired you to do some data science with some of the data it has collected about beach attendance and weather. Every year, Dreemes staffs a beach for 100 days covering the summer months. For the last ten years they have gathered data for three random variables (N=1000):

- number of daily beachgoers
- daily high temperature
- inches of daily precipitation

Each sample for each of these random variables represents the specified property (beachgoers, high temperature, precipitation) for 1 day.

From this data you generate your first plot.

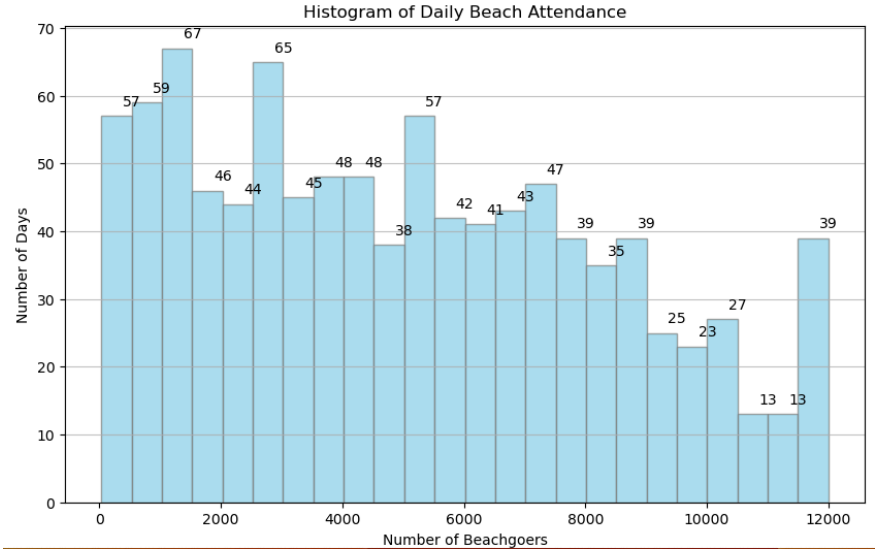

Figure 1: Histogram of daily beach attendance

Your first plot is an absolute frequency histogram of the number of daily beachgoers. The x-axis is the number of visitors, and the frequency is the number of days.

Problem 3 (10 points)

Using the histogram plot on the previous page, answer the following questions.

- a) What fraction of the summer days had under ~~250~~ 500 visitors?

  Answer: 52

- b) How many days had more than 10000 beachgoers?

  Answer: $27 + 13 + 13 + 39 = 92$

- c) What fraction of the days had more than 10000 beachgoers?

  Answer: $\frac{92}{1000} = 0.092 = 9.2\%$

  Any of these representations is acceptable.

- d) How many days have more than 15000 beachgoers?

  Answer: The distribution has no samples beyond 12000 so zero days had more than 15000 beachgoers.

- e) Is this mean the number of daily beach goers between 0 and 2000, 2000 and 4000, 4000 and 8000, or 8000 and 12000?

  Answer: 4000 and 8000. The sample mean computed from the samples used to generate this histogram is approximately 5072.

Problem 4 (10 points total, 2 points each)

You then plotted a histogram of the daily high temperatures across the 1000 days of which you have data.

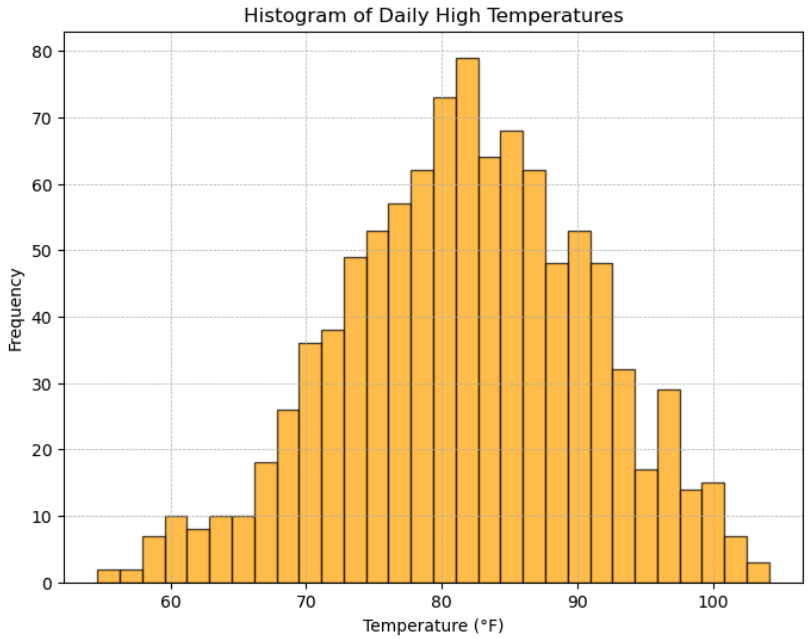

Figure 2: Histogram of daily high temperatures

- a) Which of the following distributions does this most resemble (circle one)?
  - uniform
  - exponential
  - [Gaussian]
  - bimodal

  Answer: Although it most closely approximates a Gaussian when compared to the other options, the histogram is generated from samples drawn from a scaled and shifted beta distribution. A beta distribution has two parameters that control its shape α and β. The two parameters allow us to control the skewness of the distribution. When $\alpha > \beta$, the distribution leans to the right. When $\alpha < \beta$, the distribution leans toward the left. In this case the $\alpha = 9$ and $\beta = 5$. As such, the distribution leans toward the right. A beta distribution is more appropriate for the temperature distribution because temperatures do not travel far outside a range. The temperature will

never suddenly boil. During the summer it rarely if ever freezes. I scaled and shifted the beta distribution, because a beta distribution varies between zero and one. I scaled by 78 and then shifted by 32. This ensures the beta distribution varies between 32° F and 110° F.

- b) Which of the following does this distribution exhibit (circle one)?
  - strong left skew
  - [slight left skew]
  - [no skew]
  - slight right skew
  - strong right skew

Answer: The distribution exhibits slight left skew because the beta distribution's $\alpha > \beta$. It leans to the right. A distribution exhibiting left skew has a longer and/or heavier tail extending to the left. This gives the distribution an appearance of leaning right. The tail is however not particularly heavy so I decided to also accept no skew as an answer.

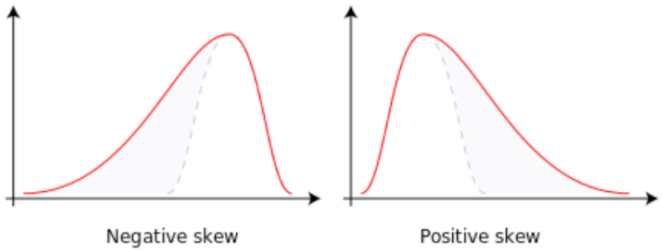

Figure 3: Negative skew results in a distribution leaning to the right, positive skew to the left.

Problem 4 (cont.)

Using the histogram of daily high temperatures on the previous page, answer the following questions.

- c) Which of the following best describes the minimum observed temperature (circle one)?
  - the mimimum is less than 0 degrees
  - the minimum is between 30 and 40 degrees
  - the minimum is between 60 and 70 degrees
  - the minimum is between 70 and 80 degrees

  Answer: I asked the students to skip this question. The underlying beta distribution has a min of 32° F, but none of the generated samples fell below the mid-50s. The mininum in this sample was 54.6° F I didn't look closely at the generated samples or I would have adjusted the choices to answer this problem.

- d) Which of the following best describes the maximum observed temperature (cirlce one)?
  - the maximum is between 90 and 95 degrees
  - the maximum is between 95 and 100 degrees
  - the maximum is between 100 and 105 degrees
  - the maximum is between 110 and 120 degrees.

  Answer: there are two histogram buckets completely above 100° F, so the maximum is definitely above 100° F. It looks like each bucket is approximately $1\frac{2}{3}$ degrees wide, and there are approximately 2 and half buckets above 100. At most the maximum could be around 105. However there is no option answering between 105 and 110 degrees, so between 100 and 105 degrees is the only reasonable answer remaining.

- e) Approximately what percentage of the days had a daily high above 100 degrees Fahrenheit?

  Answer: the key to answering this question is recognizing the use of the word "approximately." At least two histogram buckets fall completely above 100. The rightmost bucket contains 2 or maybe 3 samples. The 2nd from the right contains 6 or 7 samples. It is possible that all samples that fell in the bucket that straddles 100 were above 100° F. If this were the case then approximately 15 samples from this bucket would be above 100. If however all of the samples in this bucket are below 100 then 0 samples would contribute to the samples above 100. This establishes an upper and

lower bound on the possible values of the percentage of days above 100° F.

Let $p\%$ denote the percentage above 100.

$$100 \cdot \frac{2+6}{1000} \leq p\% \leq 100 \cdot \frac{3+7+15}{1000}$$

$$\boxed{0.8\% \leq p\% \leq 2.5\%}$$

Any percentage in this range or a specification of a range that equals or is within this range would be an acceptable answer.

- f) What best describes the sample mean of the daily high temperatures?
  - $\boxed{\text{the mean is between 80 and 85 degrees}}$
  - the mean is between 70 and 75 degrees
  - the mean is between 85 and 90 degrees
  - the mean is between 90 and 95 degrees.

Answer: The distribution exhibits slight skew to the left and thus the mean must be below the median. The peak is somewhere between 81 and 83. This means the mean probably falls somewhere between 80 and 85 degrees. It could potentially fall below 80, but there is no option for between 75 and 80 degrees, so the most likely answer is between 80 and 85 degrees. The mean computed from the samples is actually approx 81.7° F.

If we want a good estimate from the plot alone, we could perform a weighted sum of the buckets as follows:

$$\frac{1}{1000}(2 \cdot 55 + 2 \cdot 57 + 7 \cdot 58 + 10 \cdot 61 + \cdots + +7 \cdot 102 + 3 \cdot 103)$$

Next you created a table to understand the distribution of precipitation in a broad sense.

	Precipitation Range	Number of Days
0	No Rain	710
1	Light Rain (0-0.5 inches)	215
2	Moderate Rain (0.5-1.5 inches)	72
3	Heavy Rain (>1.5 inches)	3

Figure 4: Table: Dialy Precipitation (inches)

Problem 5 (5 points total)

Using the table of daily precipitation to answer the following quesitons.

- a) What percentage of the days had no precipitation? (2 points)

  Answer: $\frac{710}{1000} = \boxed{71\%}$

- b) What percentage of the days had 0.5 or more inches of precipitation? (2 points)

  Answer: $\frac{72+3}{1000} = \boxed{7.5\%}$

- c) What percentage of the days had greater than 1.5 inches of rain? (1 point)

  Answer: $\frac{3}{1000} = \boxed{0.3\%}$

Next you plotted a histogram of the precipitation in inches for rainy days. Note that the histogram omits the days which had no precipitation. We thus say that this a plot of the conditional distribution of inches of rain given that there was rain. The condition is that there was rain.

Problem 6 (5 points)

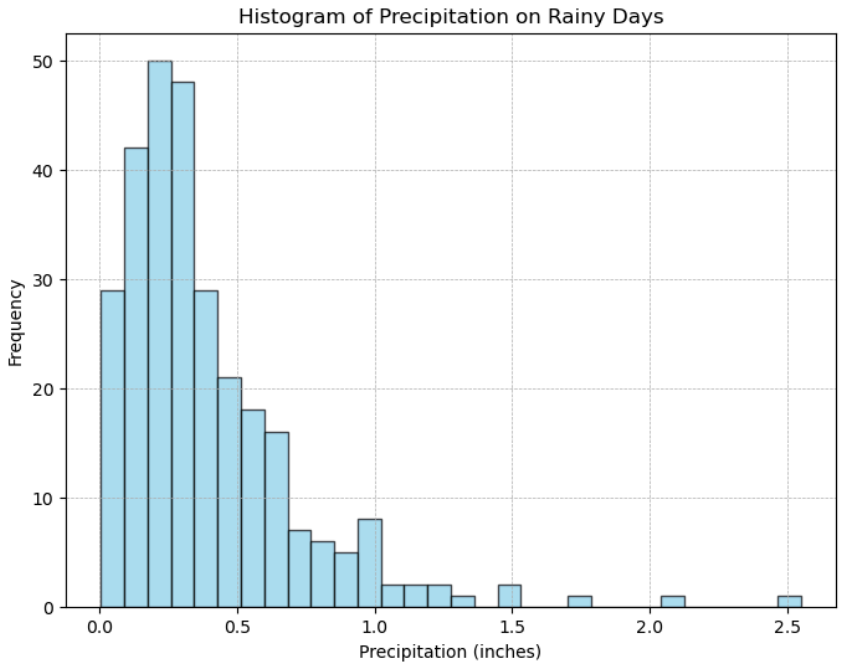

Figure 5: Histogram of daily precipitation in inches for rainy days

- a) What best describes this distribution? (2 points)
  - It has left skew.
  - It has no skew
  - It has right skew

Answer: It has a long tail to the right. This tail will pull the mean to the right of the median. If we compute skewness from the samples, the sample skewness is ≈ 3.54, which corresponds to a strong right skew.

Problem 6 (cont.)

Using the histogram from the previous page answer the following questions.

- b) Which best describes the mean precipitation on rainy days? (2 points)
  - The mean is between 0 and 0.25 inches.
  - [The mean is between 0.25 inches and 0.75 inches.]
  - The mean is between 0.75 and 1.5 inches.
  - The mean is between 1.5 and 2 inches.

  Answer: The right skewness (i.e., right tail) pulls the mean above the median. If we start adding the heights of the buckets to the left and right of the peak bucket it becomes clear that more area falls to the right of the peak so the median is somewhere above the 3rd bucket. The width of the buckets are slightly less than 0.5 inches / 5.8 buckets or about 0.9 inches per bucket. $0.9 \cdot 3 = 2.7$ If the median is greater than 2.7 and the mean is to the right of the median then the answer cannot be between 0 and 0.25 inches. There isn't nearly enough area in the tail to pull the mean above 0.75 inches, so the best answer is between 0.25 inches and 0.75 inches.

- c) How many days in the observed period had more than 2 inches of rain? (1 point)
  - Zero
  - More than 20
  - More than 5
  - [Between 0 and 5]

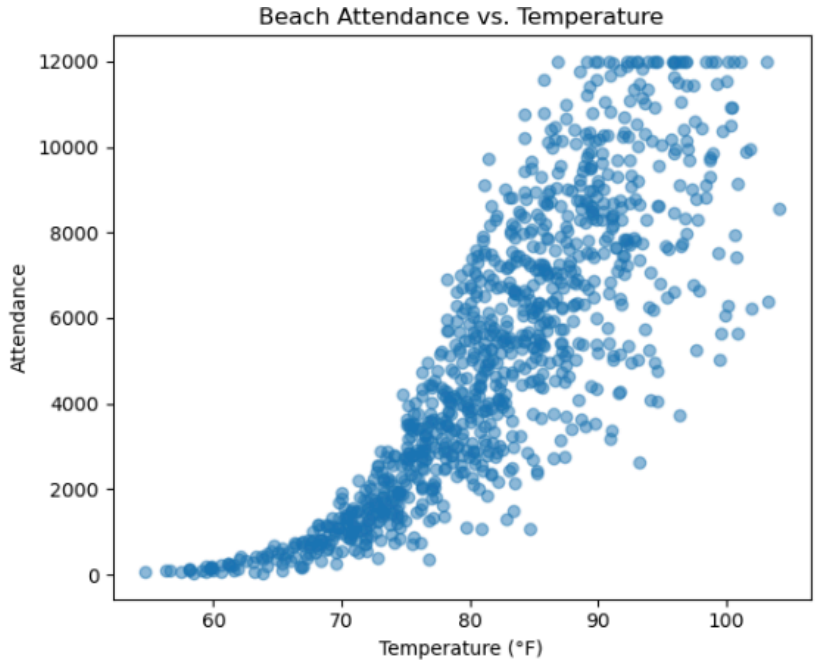

Figure 6: Number of Beachgoers vs. Temperature

Problem 7 (10 points)

Use the scatter plot of number of beachgoers vs. temperature to answer the following questions.

- a) Interpret the overall trend observed in the scatter plot between temperature and the number of beachgoers as the temperature rises. Which best describes the number of beachegoers as temperature rises? (2 points)
  - [the number of beachgoers increases with temperature]
  - the number of beachgoers decreases with temperature
  - the number of beachgoers remains the same regardless of temperature

- b) The number seems to peak or saturate at 12000. What is the most reasonable explanation? (1 point)
  - people don't like large crowds and at 12000 they all reach their limit
  - [some limitation is imposed restricting the number of people on the beach.]
  - this is just an artifact of randomness in nature

Answer: The "people don't like large crowds" may be true for most people, but it seems implausible that so many would coordinate to achieve such a sharp upper bound. It is highly likely that the distaste for large crowds would follow some distribution causing a less sharp bound on the number of attendees. Furthermore "an artifacct of randomness" also seems unlikely for the same reason as the first answer. People in large numbers do not naturally act in such tight concert. Among the options provided, the most likely explanation is that there is some restriction on the number of people on the beach. Although not offered as an explanation, another possibility worth investigating is whether there is some limitation in the way data is gathered resulting in measurement error.

- c) For this distribution would the Pearson correlation coefficient be (2 points)
  - [positive]
  - negative
  - near zero

Answer: there is clearly an upward trend in attendance as a function of temperature. The effect doesn't look linear, but a relationship only needs to trend upward to achieve a positive Pearson correlation coefficient.

Problem 8 (5 points)

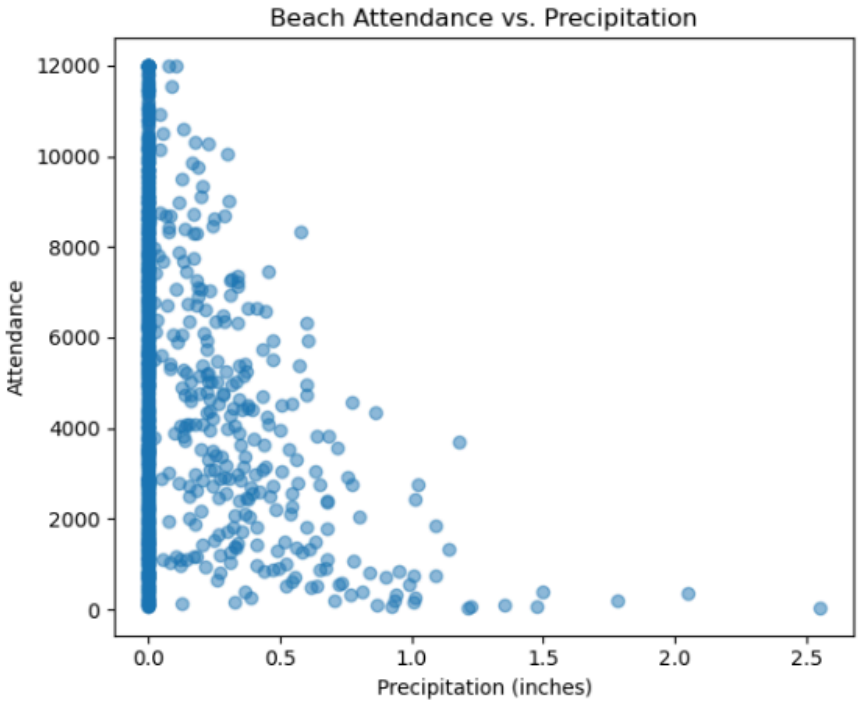

Figure 7: Number of Beachgoers vs. Precipitation

- a) What general trend does the scatter plot between daily beachgoers and precipitation suggest? (2 points)
  - Beach attendance increases with precipitation.
  - Beach attendance decreases as precipitation increases.
  - Precipitation has no visible impact on beach attendance.
  - Beach attendance peaks at moderate levels of precipitation.

  Answer: there are no samples showing heavy attendance with heavy rain. If one were to draw an average of the attendance samples as a function of the amount of attenance, without any samples showing high attendance under heavy rain, the average would come down with increasing precipitation.

- b) What best describes the Pearson correlation coefficient for this distribution? (2 points)
  - Negative

  - Near zero
  - Positive

- c) What is the reason for the near solid bar on the left-hand side of the plot? (1 point)

  Answer: It means two things, 1) there are numerous days with zero precipitation, and 2) the variability in attendance is high when there is no precipitation.

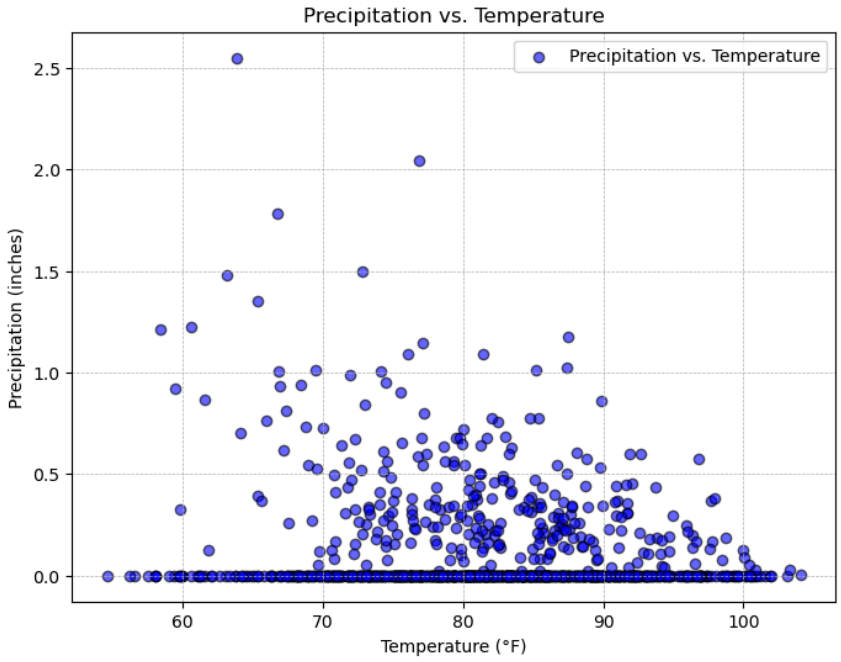

Figure 8: Precipitation versus Temperature

Problem 9 (5 points)

Use the plot in Figure 8: Precipitation versus temperature to answer the following questions.

- a) What best describes the relationship between temperature and precipitation? (2 points)
  - The amount of precipitation increases sharply with temperature
  - [The amount of precipitation diminsihes with temperature]
  - Precipitation and temperature are unrelated.
- b) What best describes the Pearson correlation coefficient for this distribution? (2 points)
  - [Negative]
  - Near zero
  - Positive

Answer: There are many samples along the bottom edge indicating that when there is no precipitation, the daily high temperature can

vary widely. However, there are no samples indicating heavy precipitation when there is high temperature. As temperature increases the instances of heavy rain seems to also diminish resulting in an overall negative correlation.

- c) What does the near solid line of samples along the bottom of the plot signify? (1 point)

  Answer: It indicates two things 1) there are many days with no precipitation, and 2) there is high variability in daily high temperature when there is no precipitation.

  Further commentary (beyond what I expect in answer to the question): when we many samples with the exact same value we have what is called a mixed random variable. It is continuous in some ranges, but it has discrete values for which a probability can be assigned. Examples of this occur everywhere in engineering. For example, a window is often completely closed or completely open, but sometimes it is opened somewhere in between. In this case we would model the system using a discrete probability of being completely open and a discrete probability of being completely closed and then use a continuous probability density function to characterize the probability of finding the window at any point in between fully open and fully closed. Rainfall is similar. There are many days with zero rainfall, but when there is precipitation, the distribution of the amount of rain would likely be best modelled with a continuous probability density function. When we have a mixed random variable, we often characterize the behavior separately for the portion with a continouus probability density function from that the behavior when the system behaves according to a discrete value (e.g., no rain).

  Although we can tell that there is significant variability in attendance when there is no rain, the fact that all the samples fall in a solid, straight line at preciptation=0 prevents us from being able to visualize the shape of the attendance distribution for non-rainy days. To do that we would probably create a histogram of the attendance specifically for non-rainy days.

Problem 10

Suppose we have the following data representing the daily number of beachgoers over eight days at Dreemes beach:

Sample Dataset S: 120, 150, 130, 170, 200, 110, 180, 140

- a) Compute the sample mean

  Answer:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{1}{8}(120 + 150 + 130 + 170 + 200 + 110 + 180 + 140)$$

$$\boxed{\overline{x} = 150}$$

- b) Compute the variance

  Answer:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$s_x^2 = \frac{1}{7}\Big((120 - 150)^2 + (150 - 150)^2 + \cdots + (140 - 150)^2\Big)$$

$$s_x^2 = \frac{1}{7}\Big(30^2 + 0^2 + 20^2 + 20^2 + 50^2 + 40^2 + 30^2 + 10^2\Big)$$

$$s_x^2 = \frac{1}{7} \cdot 6800$$

$$\boxed{s_x^2 \approx 971.43}$$

- c) Compute the sample standard deviation

  Answer:

$$s_x = \sqrt{s_x^2}$$

$$\boxed{s_x = 31.2}$$

- d) Find the minimum of the sample dataset

  Answer:

Let $x_{(i)}$ denote the i^{th} item in S when S has been sorted into ascending order. When represented in ascending order, S becomes

$$[110, 120, 130, 140, 150, 170, 180, 200]$$

Using math notation where 1 refers to the first number in a sequence.

Answer:

$$x_{(1)} = 110$$

$$\boxed{\min \ = 110}$$

- e) Find the maximum of the sample dataset

  Answer:

$$x_{(8)} = 200$$

$$\boxed{\max \ = 200}$$

- f) Find the range of the sample datasets

  Answer:

$$\text{range} = 200 - 110 = \boxed{90}$$

- g) Find the median of the sample dataset

  Answer:

  Remove an equal number of elements from the ends until either one or two element remain. If 1 element then it is the median. If two remain then taken the average of the two to compute the median.

$$[\cancel{110}, \cancel{120}, \cancel{130}, 140, 150, \cancel{170}, \cancel{180}, \cancel{200}]$$

$$\frac{140 + 150}{2} = \boxed{145}$$

- h) Find the 75th percentile of the sample dataset using linear interpolation (if necessary)

  Answer:

  Let p denote the percentile we calculating. In this case $p = 75$.

  Let r denote the rank denoting the expected index of the desire percentile. In this case the 75^{th} percentile. This may be a non-integer real number.

  $$r = \frac{p}{100} \cdot (n + 1) = \frac{75}{100} \cdot 9 = 6.75$$

  When r is a non-integer, we linearly interpolate between the numbers at indices $\lfloor r \rfloor$ and $\lceil r \rceil$.

  $$\alpha = r - \lfloor r \rfloor = 0.75$$

  Let P_p denote the value of the p^{th} percentile. In this case, we are computing P_{75}.

  $$P_{75} = x_{(6)}(1 - \alpha) + x_{(7)}\alpha$$

  $$P_{75} = 170 \cdot 0.25 + 180 \cdot 0.75 = 177.5$$

  $$\boxed{P_{75} = 177.5}$$

  If we compute percentile using the method implemented by numpy then the asnwer is 172.5. Although I prefer the 177.5 answer, I gave credit for 172.5 as well.

  When computing the percentile from a sample, we are trying to obtain the percentile of the underlying distribution. The underlying distribution is not known so at best the percentile is an approximation. People have proposed different ways to estimate the percentile with slightly different tradeoffs. In this class, I prefer the method that arrives at the answer 177.5.

- i) Find the InterQuartile Range (IQR) for the dataset.

  Answer:

  Let Q_1 denote the first quartile, i.e., the 25^{th} percentile, i.e., P_{25}.

  Let Q_3 denote the third quartile, i.e., the 75^{th} percentile, i.e., P_{75}

$$\text{IQR} = Q3 - Q1 = P_{75} - P_{25}$$

  We already know $P_{75} = 177.5$. Let's compute P_{25}.

$$r = \frac{p}{100} \cdot (n+1) = \frac{25}{100} \cdot 9 = 2.25$$

$$\alpha = r - \lfloor r \rfloor = 0.25$$

$$P_{25} = x_{(2)} \cdot (1-\alpha) + x_{(3)} \cdot \alpha = 120 \cdot 0.75 + 130 \cdot 0.25n = 122.5$$

$$\boxed{IQR = 55}$$

Problem 11

Daily High Temperature (°F): [75, 80, 85, 90, 95]

Number of Beachgoers: [200, 220, 250, 270, 300]

- a) Compute the sample mean of the daily high temperatures

  Let T be the random variable denoting a daily high temperature.

$$\bar{t} = \frac{75 + 80 + 85 + 90 + 95}{5}$$

$$\boxed{\bar{t} = 85}$$

- b) Compute the sample mean of the daily number of beachgoers

  Let B be the random variable denoting the number of daily beachgoers.

$$\bar{b} = \frac{200 + 220 + 250 + 270 + 300}{5}$$

$$\boxed{\bar{b} = 248}$$

- c) Compute the sample standard deviation of the daily high temperatures

$$s_t = \sqrt{\frac{1}{4}[75^2 + 80^2 + 85^2 + 90^2 + 95^2 - 5 \cdot 85^2]}$$

$$s_t = \sqrt{\frac{1}{4}[36375 - 36125]} = \sqrt{62.5}$$

$$\boxed{s_t \approx 7.9}$$

- c) Compute the sample standard deviation of the daily number of beachgoers

$$s_b = \sqrt{\frac{1}{4}[200^2 + 220^2 + 250^2 + 270^2 + 300^2 - 5 \cdot 248^2]}$$

$$s_b = \sqrt{\frac{1}{4}[313800 - 307520]}$$

$$s_b = \sqrt{1570}$$

$$\boxed{s_b \approx 39.6}$$

- d) Compute the sample covariance between the two sets of samples

$$s_{tb} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$s_{tb} = \frac{1}{4}\Big[(75-85)(200-248) + (80-85)(220-248) + (85-85)(250-248)$$

$$+(90-85)(270-248) + (95-85)(300-248)\Big]$$

$$s_{tb} = \frac{1}{4}\cdot 1250$$

$$\boxed{s_{tb} = 312.5}$$

- e) Compute the Pearson Corrrelation Coefficient of the two.

$$r_{tb} = \frac{s_{tb}}{s_t \cdot s_b} = \frac{312.5}{7.9 \cdot 39.6}$$

$$\boxed{r_{tb} \approx 0.998}$$