Your name: ____

ANSWER KEY for CCNY EAS 42000 / EAS A4200 Midterm 1

Friday, October 24th, 2025

Answer all questions below. This is a 100% closed exam: no electronic devices of any kind; no notes, books or other resources. Unless otherwise noted, each prompt is worth 1 point.

Please be concise and write legibly. Good luck!

1. Rank the following claims (labeled a through d) from least to most numerate:

   a. "Near-surface wind speeds are higher over the Southern Ocean than over the tropical Atlantic Ocean."

   b. "Near-surface wind speeds are, on average, 12.5 m/s over the Southern Ocean vs. 4.4 m/s over the tropical Atlantic Ocean."

   c. "Near-surface wind speeds are high over the Southern Ocean, and the reasons for that are something that scientists should investigate."

   d. "Near-surface wind speeds average 12.5±1.2 m/s over the Southern Ocean vs. 4.4±0.8 m/s over the tropical Atlantic."

   ANSWER: c, a, b, d

   - **c does not give any quantitative information ("high" compared to what?).**
   - **a at least provides a point of comparison (albeit qualitative): higher than the tropical Atlantic**
   - **b provides actual quantitative values but no uncertainty estimates**
   - **d provides quantitative values and uncertainty estimates**

2. What, if anything, is missing from all of the above options that would make them more numerate?

   ANSWER: description of *methodology*

3. You are analyzing a dataset of leaf area index (LAI). LAI is a measure of the amount of vegetation cover in a given location. The higher LAI is, the more vegetation there is. The dataset comes from retrievals made by instruments onboard a polar-orbiting satellite. Half the retrievals occur during local nighttime, and half during local daytime.

   a. It is discovered that, due to a calibration error in the sensors, all of the retrievals that were taken during local nighttime are biased low by an identical amount, specifically 0.25 m^2/m^2. Given that, how should you proceed in your analysis?

      ANSWER: add 0.25 to all nighttime retrievals

   b. Now it is discovered that, because of orbital drift of the satellite with time, the values are becoming increasingly biased toward higher values as time goes on. But the team has not yet quantified the magnitude or associated uncertainties of the resulting spurious signal. What can you say about how this would influence the dataset's timeseries?

      ANSWER: it will cause a spurious positive trend *relative* to what the trend would be otherwise.

4. Consider the following dataset of annual-average rainfall over India derived from a network of rain gauges that was in operation for five consecutive years. The values are: 3.2, 4.9, 4.8, 2.3, and 0.8 mm/day.

   a. What is the median? **ANSWER: 3.2 mm/day**

   b. What is the range? $4.9 - 0.8 = 4.1$ **mm/day**

   c. What is the sample variance? (Note that the sample mean is 3.2. Write out the equation with all the numbers plugged in, but you don't have to compute the actual value.)
   ANSWER: $N = 5$ **and** $\mu = 3.2$**, so**

$$\hat{\sigma}^2 = \frac{1}{4}((3.2\text{–}3.2)^2 + (4.9\text{–}3.2)^2 + (4.8\text{–}3.2)^2 + (2.3 - 3.2)^2 + (0.8 - 3.2)^2)$$

   d. What is the sample standard deviation, expressed in terms of the sample variance?
   ANSWER: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

   e. What is the population mean?
   ANSWER: We don't know. (Except for contrived examples, it's impossible to know the population parameters; we can only estimate them.

5. This question refers to Figure 1.

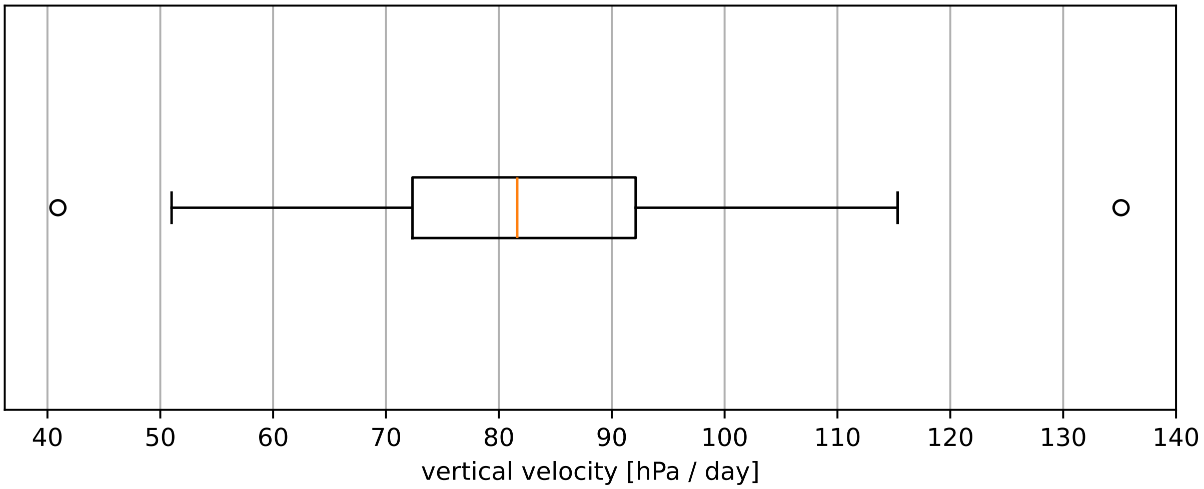

Figure 1:

a. What type of plot is this?
ANSWER: boxplot, a.k.a. box and whiskers

b. Label the median, 25th percentile, and 75th percentile on the plot. **ANSWER: the middle line near 82, the left edge of the box near 73, and the right edge of the box near 92**

c. What approximately is the range of the plotted data? **ANSWER: max minus min, approximately 136-41=95**

d. What approximately is the IQR of the plotted data? **ANSWER: 75th minus 25th percentile, approximately 92-72=20**

6. This question refers to Figure 2.

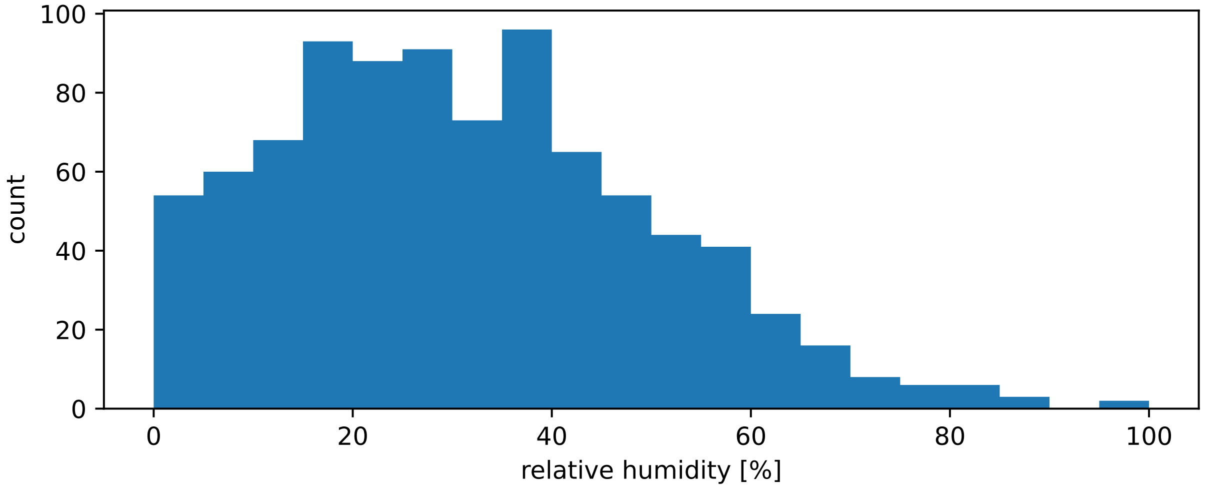

Figure 2:

a. What type of plot is this? **ANSWER: histogram**

b. Approximately how many of the values (as in to the nearest 10) in the plotted dataset are relative humidities less than 20%? **ANSWER: the four boxes below 20 have approximately 55+60+65+95=275, so call it 280**

c. How would you describe the skewness of this sample? **ANSWER: positively skewed**

d. Suppose you compute the fitted normal distribution to this sample. Would it be appropriate to directly overlay the fitted PDF on these axes? Why or why not? **ANSWER: No. This is the raw histogram, with y-axis the count, rather than the normalized histogram, such that the y-axis is count divided by bin width which is what the PDF corresponds to.**

7. This question refers to Figure 3.

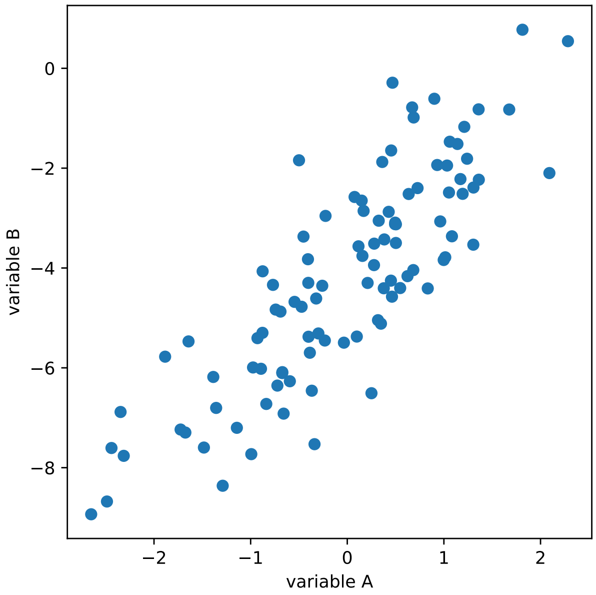

Figure 3:

a. What type of plot is this? **ANSWER: scatterplot**

b. Provide a 1-sentence description of the relationship between variables A and B. **ANSWER: pretty strong positive linear relationship**

8. If $P(E_1) + P(E_2) = 1$, what is the sample space S? **ANSWER:** $S = E_1 \cup E_2$

9. If $P(E_1) + P(E_2) = 1$, what is $E_1 \cap E_2$? **ANSWER:** $E_1 \cap E_2 = \emptyset$ **(the empty set)**

10. Describe in words what $P(A \cap B)$ means. **ANSWER: The probability of A *and* B occurring. A.k.a. the joint probability.**

11. Write the definition of $P(A \cup B)$ in terms of one or more of $P(A)$, $P(B)$, and $P(A \cap B)$. **ANSWER:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

12. (2 pts) I am interested in extreme rainfall events in New York City. To investigate it, for each year I select the single daily maximum rainfall rate in the Central Park weather station dataset that occured in that year. Is the resulting sample distribution likely to be approximately Gaussian? Briefly explain why or why not.

    ANSWER: No, this is a block maximum, which typically follow the *GEV* distribution rather than a Gaussian.

13. A student is analyzing a timeseries of dissolved organic carbon measurements taken from an instrument deployed in the Long Island Sound. Each measurement is spaced 10 minutes apart. The student computes averages over each calendar day from these samples, over a study period of 90 days. There is no strong diurnal (i.e. day/night) cycle or seasonal cycle in this variable.

    a. (2 pts) Based on this information, what, if anything, can we expect about the population of these daily averages? Explain your reasoning. (1–2 sentences)

       ANSWER: It will be Gaussian. Each day is an average of a large number ($24 \times 6 = 144$) of samples, and the small diurnal and seasonal cycles means that the individual measurements can be considered iid (independent and identically distributed). These are the conditions for the Central Limit Theorem to hold.

    b. (2 pts) Next, the student computes the sample mean to be 20 mg/L and the sample variance to be 100 (mg/L)^2. Sketch likely graphs of the corresponding sample PDF and CDF.

    ANSWER: Gaussian with mean at 20 mg/L and standard deviation 10 mg/L.

14. (2 pts) Consider a random number generator that returns one of the following four numbers selected at random: 5, 7.5, 10, and 12. The probabilities of each number being selected are, respectively, 0.2, 0.3, 0.1, and 0.4. Sketch the graphs, or write out mathematically, of the probability mass function and cumulative distribution function.

ANSWER: PMF(x)=0.2, 0.3, 0.1, and 0.4 for the numbers 5, 7.5, 10, and 12, respectively. CDF(x)=0.2, 0.5, 0.6, and 1, respectively

15. If $f(x)$ is a probability density function, then:

    a. What is $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$? **ANSWER: 1 (all PDFs satisfy this property: the integral over all values is exactly 1.**

    b. What is the corresponding cumulative distribution function $F(x)$? **ANSWER: By definition,** $F(x) = \int_{-\infty}^{x} f(x)$

16. An atmospheric chemist collects measurements of ozone (O_3) concentrations, measured in parts per billion (ppb), at a field station located in midtown Manhattan. She collects 10 data points, each one separated by 10 minutes, starting at 12 noon and ending at 1:30pm. Their mean is 23.5 ppb.

    a. What is the sample size? **ANSWER: 10**

    b. What is the sample mean? **ANSWER: 23.5 ppb**

    c. What is the population mean? **ANSWER: It's impossible to know**

17. Use Table 1 (on the next page) to compute the following empirical probabilities.

    ANSWER: Notice: because 3 days have missing days, those should *not* be included in the denominator of your calculations.

    a. $P(\text{min. temp} < 0°\text{F})$ **ANSWER: 7/28=1/4**

    b. $P(\text{max temp} > 20°\text{F} \,|\, \text{min. temp} < 0°\text{F})$ **ANSWER: 5/7**

    c. $P(\text{max temp} > 40°\text{F})$ **ANSWER: 2/28=1/14**

    d. $P(\text{precip} > 1.5'')$ **ANSWER: 0/28=0**

day	precip (in)	max T (°F)	min T (°F)
1	0.00	33	19
2	0.07	32	25
3	1.11	30	22
4	0.00	29	-1
5	0.00	25	4
6	0.00	30	14
7	0.00	37	21
8	0.04	37	22
9	0.02	29	23
10			
11			
12			
13	0.18	33	29
14	0.02	34	15
15	0.02	53	29
16	0.00	45	24
17	0.00	25	0
18	0.00	28	2
19	0.00	32	26
20	0.45	27	17
21	0.00	26	19
22	0.00	28	9
23	0.70	24	20
24	0.00	26	-6
25	0.00	9	-13
26	0.00	22	-13
27	0.00	17	-11
28	0.00	26	-4
29	0.01	27	-4
30	0.03	30	11
31	0.05	34	23

Table 1: Daily weather conditions in Ithaca, NY, measured in January 1987. The columns are, from left to right, the day of the month, precipitation (inches), maximum temperature (°F), and minimum temperature (°F). Blank entries indicate missing data.

Blank pages for extended answers and/or scratch work: If you need extra room for any of your answers, put them here. Make sure you identify which question each one corresponds to. You can also use this for scratch work, which will not be graded. Just make clear which parts are scratch work and which parts are the actual answer extensions you want graded.