

2026



AP[®] Statistics

Free-Response Questions

STATISTICS

SECTION II

TIME – 1 HOUR AND 30 MINUTES

Directions:

Section II has 6 free-response questions and lasts 1 hour and 30 minutes.

You may use the available paper for scratch work and planning, but only work written in the free-response booklet will be scored. Any work done on scratch paper will not be scored. Label parts (e.g., A, B, C) and sub-parts (e.g., i, ii, iii) as needed. Use a pencil or a pen with black or dark blue ink to write your responses.

Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations. Correct answers without supporting work may not receive credit.

You may use a handheld calculator in this section or the calculator available in this application.

Reference information, including lists of formulas and tables, can be used throughout the exam. A digital version is available in this application.

You may pace yourself as you answer the questions in this section, or you may use these optional timing recommendations:

It is suggested that you spend about 1 hour and 5 minutes on questions 1 through 5 and about 25 minutes on question 6.

You can go back and forth between questions in this section until time expires. The clock will turn red when 5 minutes remain—the proctor will not give you any time updates or warnings.

Note: This exam was originally administered digitally. It is presented here in a format optimized for teacher and student use in the classroom.

During the AP Exam administration, students have access to reference information. To see the reference information for this course, please visit AP Central:
<https://apcentral.collegeboard.org/exam-administration-ordering-scores/administering-exams/subject-specific/reference-information>

1. A goat farmer raises two breeds of goats, Breed H and Breed J. He wants to compare the weights of the two breeds and takes independent random samples of 14 goats from each breed. The weight, in pounds, of each goat is measured. The weights of the 14 Breed H goats are recorded.

Weights of Breed H Goats (pounds)

48
48
55
56
56
57
62
66
72
72
72
73
80
80

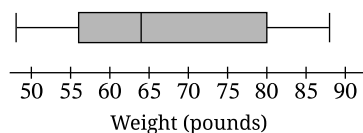
Part A

Use the data in the list to determine the five-number summary for the sample of Breed H goats.

Part B

The distribution of weight for Breed J goats is displayed in the boxplot.

Weights of Breed J Goats (pounds)



Use the five-number summary from part A and the boxplot of Breed J to compare the center and variability for the distribution of weight for the Breed H goats and the distribution of weight for the Breed J goats, in context.

Part C

A stem-and-leaf plot is another way to display the weights of goats. The distribution of weight for the Breed H goats is displayed in the given stem-and-leaf plot.

Weights of Breed H Goats (pounds)

4		8 8
5		5 6 6 7
6		2 6
7		2 2 2 3
8		0 0

Key: 4|8 = 48 pounds

Consider the five-number summary from part A and the stem-and-leaf plot.

- If a boxplot were created from the five-number summary found in part A, what characteristic of the shape of the distribution of weight for the Breed H goats would be apparent from the stem-and-leaf plot but not from the boxplot?
- Explain why a boxplot would not display the characteristic of the shape of the distribution identified in part C (i).

2. Holly, a **botanist**, read an online report that stated that adding coffee grounds to the soil of rosebushes will produce more roses. Holly decides to conduct an experiment to investigate this claim. She grows 30 rosebushes in a greenhouse with a controlled environment so the conditions will be the same for all rosebushes.

When the rosebushes are a month old, she randomly assigns 15 rosebushes to have one-half cup of coffee grounds added to their soil weekly. The other 15 rosebushes will not have coffee grounds added to their soil.

After three months, she counts the number of roses on each rosebush.

botanist: a scientist who studies the biology of plants

Part A

Based on the information provided about Holly's experiment, identify each of the following.

- i. Treatments
- ii. Experimental units
- iii. Response variable

Part B

In the context of Holly's experiment, describe how the treatments can be randomly assigned to the experimental units so that each treatment has the same number of units.

Part C

After the experiment, Holly determines that her results were statistically significant at the $\alpha = 0.05$ level of significance. Explain the meaning of "statistically significant" in this context.

3. A certain sports team has its team song performed by a different musician at the beginning of each of its games. The time it takes for the team song to be performed by different musicians can be modeled using a normal distribution with mean $\mu = 109$ seconds and standard deviation $\sigma = 16$ seconds. All performances of the team song are independent.

Part A

What is the probability that a randomly selected performance of the team song will take longer than 120 seconds? Show your work.

Part B

Ten performances of the team song will be randomly selected. Let the random variable X represent the number of games at which the performance of the team song takes longer than 120 seconds. Find $P(X \geq 3)$. Show your work.

Part C

Ben will attend all of the games for this sports team. Let the random variable Y represent the number of games Ben will attend until a performance of the team song takes longer than 120 seconds.

- i. Calculate the mean of Y . Show your work.
- ii. Calculate the standard deviation of Y . Show your work.

Part D

Interpret the standard deviation calculated in part C (ii) in context.

4. A farmer wants to know whether there is a difference in the mean number of oranges that grow on trees fertilized with the current fertilizer, Brand C, or a new fertilizer, Brand N. The farmer randomly assigns 58 trees to be treated with Brand C and 58 trees to be treated with Brand N. The number of oranges from each tree is recorded. The summary statistics by fertilizer brand are shown in the table.

Summary Statistics for the Number of Oranges by Fertilizer Brand

	n	Mean	Standard Deviation
Brand C	58	141	15
Brand N	58	148	19

At the $\alpha = 0.05$ significance level, do the data provide convincing statistical evidence that the mean number of oranges on trees fertilized with Brand C is different from the mean number of oranges on trees fertilized with Brand N for all trees similar to those in the study?

5. A researcher is investigating whether there is an association among professional athletes between age-group (in years) and type of sport played (basketball, football, baseball). The age-group and type of sport played for all 4,193 professional athletes in these sports for a recent year are given in the two-way table.

Age-Group by Type of Sport Played

Age (years)	Basketball	Football	Baseball	Total
Age < 25	232	807	259	1,298
$25 \leq \text{Age} < 30$	175	1,326	620	2,121
$30 \leq \text{Age} < 35$	90	287	276	653
$35 \leq \text{Age}$	19	41	61	121
Total	516	2,461	1,216	4,193

Part A

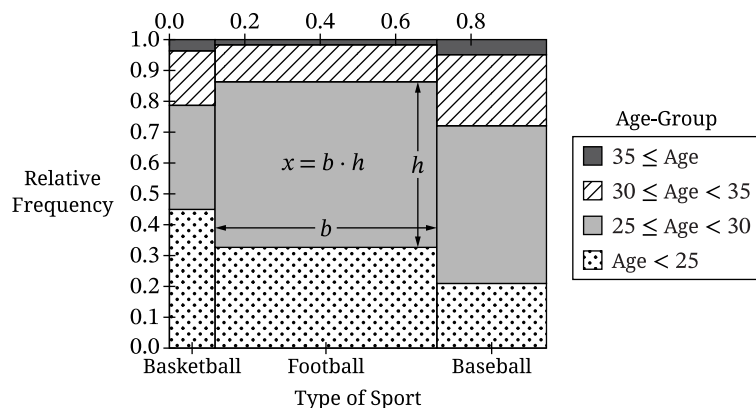
Consider the two-way table.

- What is the probability a randomly selected professional athlete is a football player? Show your work.
- What is the probability a randomly selected professional athlete is in the age-group of " $25 \leq \text{Age} < 30$ " given they are a football player? Show your work.

Part B

A mosaic plot was constructed using the information in the two-way table. Use the mosaic plot to answer the following questions.

Relative Frequency of Age-Group by Type of Sport Played



- The b and h displayed in the mosaic plot each represent a probability calculated in part A. Which probability does b correspond to: the probability calculated in part A (i) or the probability calculated in part A (ii)?
- What probability does the x displayed in the mosaic plot represent in context?

Part C

Use the information to answer the following questions.

- Are the events “Baseball” and “35 ≤ Age” mutually exclusive events? Explain.
- Are the events “Baseball” and “35 ≤ Age” independent events? Show your work.

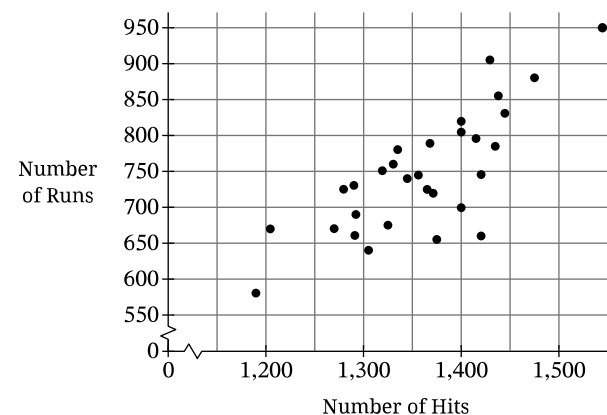
Part D

Consider the data of all professional athletes from the two-way table. Determine if it is appropriate to carry out a chi-square test for independence to investigate whether there is an association between age-group and sport played using these data. Explain your answer.

The following information applies to parts A and B.

- To win a game, a baseball team needs to score runs. To score runs, players need to get on base. The primary way players get on base is by hitting the ball. Figure 1 shows the relationship between the number of hits and the number of runs for 30 randomly selected professional baseball teams.

Figure 1. Scatterplot of Number of Runs by Number of Hits

**Part A**

Consider the scatterplot in Figure 1.

- Describe the relationship between the number of hits and the number of runs for these baseball teams, in context.
- Based on the data from the sample of 30 teams, an equation of the least-squares regression line model for predicting the number of runs from the number of hits is as follows.

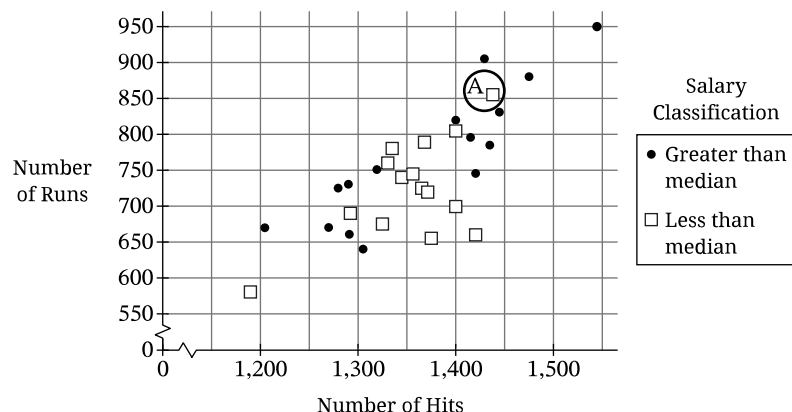
$$\text{Predicted number of runs} = -372.2 + 0.823(\text{number of hits})$$

Using the given regression equation, calculate the point estimate for the predicted number of runs a team would score if they were to achieve their goal of 1,250 hits. Show your work.

Part B

Each team has a total salary equal to the sum of the salaries of all players on the team. The median total salary for the sample of 30 teams is \$160 million. In Figure 2, points with dots represent teams with a total salary greater than the median, and points with squares represent teams with a total salary less than the median.

Figure 2. Scatterplot of Number of Runs by Number of Hits and Salary Classification



Consider the scatterplot in Figure 2.

- Compare the team represented by the point that is circled and labeled A with the other teams that have the same total salary classification.
- For each salary classification, consider the linear relationship between the number of hits and the number of runs. For teams with a total salary greater than the median, is the strength of the linear relationship stronger than, weaker than, or similar to the strength of the linear relationship for teams with a total salary less than the median? Explain.

Part C

Different types of intervals can be used when working with linear regression models. One type of interval is a confidence interval for the slope of the least-squares regression line. Two other types of intervals in the context of this problem are as follows:

- A confidence interval that estimates the mean number of runs for all teams with a specific number of hits.
- A prediction interval that predicts the number of runs for a single team with a specific number of hits.

All three intervals use the same basic formula:

$$\text{Point Estimate} \pm t^*(\text{standard error}),$$

where critical value t^* comes from a t -distribution that has $n - 2$ degrees of freedom. The same critical value is used when finding each of these intervals with the same level of confidence.

Consider the point estimate for the predicted number of runs found in part A (ii) for the team whose goal is to achieve 1,250 hits next year. Recall that this value was calculated using the linear regression model based on the data from the sample of 30 teams.

- What is the critical value for a 95% confidence level that would be used for the confidence interval for the mean number of runs as well as for the prediction interval for the number of runs? Indicate the answer to two decimal places.
- The standard error for the confidence interval for the mean number of runs is 17.48. Assuming the conditions for inference are met, calculate the 95% confidence interval for the mean number of runs for all teams with 1,250 hits. Show your work.
- The standard error for the prediction interval for the number of runs is 56.78. Assuming the conditions for inference are met, calculate the 95% prediction interval for the number of runs for a single team with 1,250 hits. Show your work.

Part D

- i. Would a distribution of sample means be expected to have more variability or less variability than a distribution of individual observations? Explain.
- ii. The standard error used in the confidence interval that estimates the mean number of runs for all teams with x hits can be found using the following formula.

$$\sqrt{s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

The standard error used in the prediction interval that predicts the number of runs for a single team with x hits can be found using the following formula.

$$\sqrt{s^2 + s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

In both standard error formulas, s is the standard deviation of the residuals, n is the sample size, and $\bar{x}$ is the sample mean number of hits. Based on the answer from part D (i) and the standard error formulas for the confidence interval and the prediction interval, explain why the prediction interval calculated in part C (iii) is wider than the confidence interval calculated in part C (ii).

STOP
END OF EXAM