http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

AP Statistics Exam Tips for Students

The following exam tips were written by Sanderson Smith, retired, and Daren Starnes of The Lawrenceville School, Lawrenceville, New Jersey, and are used with permission.

The Exam Itself

To maximize your score on the AP Statistics Exam, you first need to know how the exam is organized and how it will be scored.

The AP Statistics Exam consists of two separate sections:

Section I	40 Multiple-Choice questions	90 minutes	counts 50 percent of exam score

SCORING:
1 point for each correct answer
0 points for each question left blank

Section II	Free-Response questions	90 minutes	counts 50 percent of exam score	Questions are designed to test your statistical reasoning and your communication skills.

SCORING:
Five open-ended problems @ 13 minutes; each counts 15 percent of free-response score
One investigative task @ 25 minutes; counts 25 percent of free-response score

Each free-response question is scored on a 0 to 4 scale. General descriptors for each of the scores are:

4	Complete Response	NO statistical errors and clear communication
3	Substantial Response	Minor statistical error/omission or fuzzy communication
2	Developing Response	Important statistical error/omission or lousy communication
1	Minimal Response	A "glimmer" of statistical knowledge related to the problem
0	Inadequate Response	No glimmer; statistically dangerous to himself and others

Your work is graded **holistically**, meaning that your entire response to a problem is considered before a score is assigned.

Exam preparation begins on the first day of your AP Statistics class. Keep in mind the following advice throughout the year:

- Read your statistics book. Most AP Exam questions start with a paragraph that describes the context of the problem. You need to be able to pick out important statistical cues. The only way you will learn to do that is through hands-on experience.
- Practice writing about your statistical thinking. Your success on the AP Exam depends on how well you explain your reasoning.
- Work as many problems as you can in the weeks leading up to the exam. Your biggest challenge will be determining what statistical technique to use on each question.

On the night before the exam

- Get a good night's sleep.

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

- Make sure your calculator is functioning properly. Insert new batteries, and make sure all systems are "go." Bring a spare calculator if possible.

During the Exam

General Advice

Relax, and take time to think! Remember that everyone else taking the exam is in a situation identical to yours. Realize that the problems will probably look considerably more complicated than those you have encountered in other math courses. That's because a statistics course is, necessarily, a "wordy" course.

Read each question carefully before you begin working. This is especially important for problems with multiple parts or lengthy introductions. Suggestion: Underline key words and phrases as you read the questions.

Look at graphs and displays carefully. For graphs, note carefully what is represented on the axes, and be aware of number scale. Some questions that provide tables of numbers and graphs relating to the numbers can be answered simply by "reading" the graphs.

About graphing calculator use: Your graphing calculator is meant to be a tool, to be used sparingly on some exam questions. Your brain is meant to be your primary tool.

On multiple-choice questions:

- Examine the question carefully. What statistical topic is being tested? What is the purpose of the question?
- Read carefully. Underline key words and phrases. After deciding on an answer, glance at the underlined words and phrases to make sure you haven't made a careless mistake or an incorrect assumption.
- You don't have to answer all of the questions to get a good overall score.
- If an answer choice seems "obvious," think about it. If it's so obvious to you, it's probably obvious to others, and chances are good that it is not the correct response. For example, suppose one set of test scores has a mean of 80, and another set of scores on the same test has a mean of 90. If the two sets are combined, what is the mean of the combined scores. The "obvious" answer is 85 (and will certainly appear among the answer choices), but you, as an intelligent statistics student, realize that 85 is not necessarily the correct response.

On free-response questions:

- Do not feel pressured to work the free-response problems in a linear fashion, for example, 1, 2, 3, 4, 5, 6. Read all of the problems before you begin. Question 1 is meant to be straightforward, so you may want to start with it. Then move to another problem that you feel confident about. Whatever you do, don't run out of time before you get to Question 6. This Investigative Task counts almost twice as much as any other question.
- Read each question carefully, sentence by sentence, and underline key words or phrases.
- Decide what statistical concept/idea is being tested. This will help you choose a proper approach to solving the problem.
- You don't have to answer a free-response question in paragraph form. Sometimes an organized set of bullet points or an algebraic process is preferable.
- Answer each question in context.

Specific Advice on Free-Response Questions

On problems where you have to produce a graph:

- Label and scale your axes! Do not copy a calculator screen verbatim onto the test.
- Don't refer to a graph on your calculator that you haven't drawn. Transfer it to the exam paper. This is part of your burden of good communication.

Communicate your thinking clearly.

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

- Organize your thoughts before you write, just as you would for an English paper.
- Write neatly.
- Write efficiently. Say what needs to be said, and move on. Don't ramble.
- The burden of communication is on you. Don't leave it to the reader to make inferences.
- Don't contradict yourself.
- Avoid bringing your personal ideas and philosophical insights into your response.
- When you finish writing your answer, look back. Does the answer make sense? Did you address the context of the problem?

About graphing calculator use:

- Don't waste time punching numbers into your calculator unless you're sure it is necessary. Entering lists of numbers into a calculator can be time-consuming, and certainly doesn't represent a display of statistical intelligence.
- Do not write directions for calculator button-pushing on the exam!
- Avoid calculator syntax, such as *normalcdf* or *1-PropZTest.*

Follow directions. If a problem asks you to "explain" or "justify," then be sure to do so.

- Don't "cast a wide net" by writing down everything you know, because you will be graded on everything you write. If part of your answer is wrong, you will be penalized.
- Don't give parallel solutions. Decide on the best path for your answer, and follow it through to the logical conclusion. Providing multiple solutions to a single question is generally not to your advantage. You will be graded on the lesser of the two solutions. Put another way, if one of your solutions is correct and another is incorrect, your response will be scored "incorrect."

The amount of space provided on the free-response questions does not necessarily indicate how much you should write.

If you cannot get an answer to part of a question, make up a plausible answer to use in the remaining parts of the problem.

Content-Specific Tips

I. Exploring Data

When you analyze one-variable data, always discuss shape, center, and spread.

Look for patterns in the data, and then for deviations from those patterns.

Don't confuse *median* and *mean*. They are both measures of center, but for a given data set, they may differ by a considerable amount.

(a) If distribution is skewed right, then mean is greater than median.
(b) If distribution is skewed left, then mean is less than median.
Mean > median is not sufficient to show that a distribution is skewed right.
Mean < median is not sufficient to show that a distribution is skewed left. **Don't confuse *standard deviation and variance***. Remember that standard deviation units are the same as the data units, while variance is measured in square units.

Know how transformations of a data set affect summary statistics.
(a) Adding (or subtracting) the same positive number *k,* to (from) each element in a data set increases (decreases) the mean and median by *k.* The standard deviation and IQR do not change.
(b) Multiplying all numbers in a data set by a constant *k* multiplies the mean, median, IQR, and standard deviation by *k.* For instance, if you multiply all members of a data set by four, then the new set has a standard deviation that is four times larger than that of the original data set, but a variance that is 16 times the original variance.
Simple examples:

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

Original data set	Mean	St. Dev.	Variance	Median	IQR	Range
{1,2,3,4,5}	3	1.414	2	3	3	4

Add 7 to each element of the original data set:

New data set	Mean	St. Dev.	Variance	Median	IQR	Range
{8,9,10,11,12}	10	1.414	2	10	3	4

Multiply each element of the original data set by four:

New data set	Mean	St. Dev.	Variance	Median	IQR	Range
{4,8,12,16,20}	12	5.6569	31	12	12	16

Multiply elements of the original data set by four, then add seven:

New data set	Mean	St. Dev.	Variance	Median	IQR	Range
{11,15,19,23,27}	19	5.6569	32	19	12	16

When commenting on shape:

- Symmetric is not the same as "equally" or "uniformly" distributed.
- Do not say that a distribution "is normal" just because it looks symmetric and unimodal.

Treat the word "normal" as a "four-letter word." You should only use it if you are really sure that it's appropriate in the given situation.

When describing a scatterplot:

- Comment on the direction, shape, and strength of the relationship.
- Look for patterns in the data, and then for deviations from those patterns.

A correlation coefficient near 0 doesn't necessarily mean there are no meaningful relationships between the two variables. Consider the following data points:

X	2	3	4	5	6	7	8	9	10	11	12
Y	6	30	8	50	10	70	12	90	14	110	16

In this case, $r = .38$, indicating fairly weak correlation, but a scatterplot displays something quite interesting. Moral of the story: Always plot your data.

Don't confuse correlation coefficient and slope of least-squares regression line.

- A slope close to 1 or -1 doesn't mean strong correlation.
- An *r* value close to 1 or -1 doesn't mean the slope of the linear regression line is close to 1 or -1.
- The relationship between *b* (slope of regression line) and *r* (coefficient of correlation) is

$$b = r \cdot \frac{s_y}{s_x}$$

This is on the formula sheet provided with the exam.

- Remember that $r^2 > 0$ doesn't mean $r > 0$. For instance, if $r^2 = 0.81$, then $r = 0.9$ or $r = -0.9$.

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

You should know difference between a scatter plot and a residual plot.

For a residual plot, be sure to comment on:

- The balance of positive and negative residuals
- The size of the residuals relative to the corresponding y-values
- Whether the residuals appear to be randomly distributed

Given a least squares regression line, you should be able to correctly interpret the slope and y-intercept in the context of the problem.

Remember properties of the least-squares regression line:

- Contains the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of the *x*-values and $\bar{y}$ is the mean of the *y*-values.
- Minimizes the sum of the squared residuals (vertical deviations from the LSRL)

Residual = (**actual y-value** of data point) - (**predicted y-value** for that point from the LSRL)

Realize that logarithmic transformations can be practical and useful. Taking logs cuts down the magnitude of numbers. Also, if there is an exponential relationship between *x* and *y* ($y=ab^x$), then a scatterplot of the points {(*x,log y*)} has a linear pattern.

Example:

x	*y*	*log y*
1	24	1.3802
2	192	2.2833
3	1,536	3.1864
4	12,188	4.0859
7	6,290,000	6.7987
8	49,900,000	7.6981

An **exponential fit** to (x,y) on the TI-83 yields $y = 3.002(7.993^x)$, with $r = 0.9999$. When $x = 9$, this model predicts y = 399,901,449.2.

A **linear fit** to (*x,log y*) on the TI-83 yields $\log y = 0.477395 + 0.9027286x$, with $r = .9999$. If $x = 9$, then $\log y = 0.477395 + 0.9027286(9) = 8.601952978$. Hence $y = 10^{8.601952978} = 399,901,449.2$.

If the relationship between *x* and *y* is described by a power function ($y=ax^b$), then a scatterplot of (log*x*, log *y*) will have a linear pattern.

Example:

x	y	log x	log y
1	8	0	.90309
2	64	.30103	1.8062
3	216	.47712	2.3345
4	512	.60206	2.7093
7	2744	.8451	3.4384
8	4096	.90309	3.6124

A **power fit** to (*x,y*) on the TI-83 yields $y=8x^3$ with $r=1$. When $x=9$, this model predicts $y=8(9)^3$ =5832.

A **linear fit** to (log *x*, log *y*) on the TI-83 yields log $y = .90309 + 3 \log x$ with $r = 1$. When $x = 9$, this model predicts log $y = .90309 + 3 \log(9) = 3.76582$
Hence, $y = 10^{3.76582} = 5832$.

II. Surveys, observational studies, and experiments

Know what is required for a sample to be a **simple random sample** (SRS). If each individual in the population has an equal probability of being chosen for a sample, it doesn't follow that the sample is an SRS. Consider a class of six boys and six girls. I want to randomly pick a committee of two students from this group. I decide to flip a coin. If "heads," I will choose two girls by a random process. If "tails," I will choose two boys by a random process. Now, each student has an equal probability (1/6) of being chosen for the committee. However, the two students are not an SRS of size two picked from members of the class. Why not? Because this selection process does not allow for a committee consisting of one boy and one girl. To have an SRS of size two from the class, each group of two students would have to have an equal probability of being chosen as the committee.

SRS refers to how you obtain your sample; random allocation is what you use in an experiment to assign subjects to treatment groups. They are not synonyms.

Well-designed experiments satisfy the principles of **control, randomization, and replication**.

- Control for the effects of lurking variables by comparing several treatments in the same environment. **Note:** Control is not synonymous with "control group."
- Randomization refers to the random allocation of subjects to treatment groups, and not to the selection of subjects for the experiment. Randomization is an attempt to "even out" the effects of lurking variables across the treatment groups. Randomization helps avoid bias.
- Replication means using a large enough number of subjects to reduce chance variation in a study. **Note:** In science, replication often means, "do the experiment again."

Distinguish the language of surveys from the language of experiments. Stratifying:sampling::Blocking:experiment

It is not enough to memorize the terminology related to surveys, observational studies, and experiments. You must be able to **apply** the terminology in context. For example:

Blocking refers to a deliberate grouping of subjects in an experiment based on a characteristic (such as gender, cholesterol level, race, or age) that you suspect will affect responses to treatments in a systematic way. After blocking, you should randomly assign subjects to treatments within the blocks. Blocking reduces unwanted variability.

An experiment is **double blind** if neither the subjects nor the experimenters know who is receiving what treatment. A third party can keep track of this information.

Suppose that subjects in an observational study who eat an apple a day get significantly fewer cavities than subjects who eat less than one apple each week. A possible **confounding variable** is overall diet. Members of the apple-a-day group may tend to eat fewer sweets, while those in the non-apple-eating group may turn to sweets as a regular alternative. Since different diets could contribute to the disparity in cavities between the two groups, we cannot say that eating an apple each day causes a reduction in cavities.

III. Anticipating patterns: probability, simulations, and random variables

You need to be able to describe how you will perform a simulation in addition to actually doing it.

- Create a correspondence between random numbers and outcomes.
- Explain how you will obtain the random numbers (e.g., move across the rows of the random digits table, examining pairs of digits), and how you will know when to stop.
- Make sure you understand the purpose of the simulation -- counting the number of trials until you achieve "success" or counting the number of "successes" or some other criterion.

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

- Are you drawing numbers with or without replacement? Be sure to mention this in your description of the simulation and to perform the simulation accordingly.

If you're not sure how to approach a probability problem on the AP Exam, see if you can design a simulation to get an approximate answer.

Independent events are not the same as *mutually exclusive* (disjoint) events.

Two events, A and B, are *independent* if the occurrence or non-occurrence of one of the events has no effect on the probability that the other event occurs.

Events A and B are *mutually exclusive* if they cannot happen simultaneously.

Example: Roll two fair six-sided dice. Let A = the sum of the numbers showing is 7, B = the second die shows a 6, and C = the sum of the numbers showing is 3.

By making a table of the 36 possible outcomes of rolling two six-sided dice, you will find that P(A) = 1/6, P(B) = 1/6, and P(C) = 2/36.

- Events A and B are independent. Suppose you are told that the sum of the numbers showing is 7. Then the only possible outcomes are {(1,6), (2,5), (3,4), (4,3), (5,2), and (6,1)}. The probability that event B occurs (second die shows a 6) is now 1/6. This new piece of information did not change the likelihood that event B would happen. Let's reverse the situation. Suppose you were told that the second die showed a 6. There are only six possible outcomes: {(1,6), (2,6), (3,6), (4,6), (5,6), and (6,6)}. The probability that the sum is 7 remains 1/6. Knowing that event B occurred did not affect the probability that event A occurs.
- Events A and B are not disjoint. Both can occur at the same time.
- Events B and C are mutually exclusive (disjoint). If the second die shows a 6, then the sum cannot be 3. Can you show that events B and C are not independent?

In symbols, $P(A|B) = P(A)$ and $P(B|A) = P(B)$ if and only if events A and B are independent.

Only the "general" probability rules are provided on the AP exam.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Recognize a *discrete random variable* setting when it arises. Be prepared to calculate its mean (expected value) and standard deviation.

Example:

Let X = the number of heads obtained when five fair coins are tossed.

Value of x	0	1	2	3	4	5
Probability	1/32 = 0.03125	5/32 = 0.15625	10/32 = 0.3125	10/32 = 0.3125	5/32 = 0.15625	1/32 = 0.03125

http://apcentral.collegeboard.com/apc/members/exam/exam_tips/4073.html

$E(X) = \mu_x$
$= 0(.03125) + 1(.15625) + 2(.3125) + 3(.3125) + 4(.15625) + 5(.03125) = 2.5.$

$Var(X) = \sigma_x{}^2$
$= .03125(0-2.5)^2 + .15625(1-2.5)^2 + .3125(2-2.5)^2 + .3125(3-2.5)^2$
$+ .15625(4-2.5)^2 + .03125(5-2.5)^2 = 1.25.$

So $\sigma_x = \sqrt{1.25} = 1.118.$

You need to be able to work with transformations and combinations of random variables.

- For <u>any</u> random variables X and Y:
  $\mu_{a+bX} = a + b\mu_X$ and $\sigma^2{}_{a+bX} = b^2\sigma^2{}_X$
  $\mu_{X+Y} = \mu_X + \mu_Y$
  $\mu_{X-Y} = \mu_X - \mu_Y$
- For <u>independent</u> random variables X and Y:

  $\sigma^2_{X+Y} = \sigma^2_x + \sigma^2_y$

Expected value ($E(X)$ or μ_X) does not have to be a whole number.

Recognize a *binomial situation* when it arises.

The four requirements for a chance phenomenon to be a binomial situation are:

1. There are a fixed number of trials.
2. On each trial, there are two possible outcomes that can be labeled "success" and "failure."
3. The probability of a "success" on each trial is constant.
4. The trials are independent.

Example: Consider rolling a fair die 10 times. There are 10 trials. Rolling a 6 constitutes a "success," while rolling any other number represents a "failure." The probability of obtaining a 6 on any roll is 1/6, and the outcomes of successive trials are independent.

Using the TI-83, the probability of getting exactly three sixes is $({}_{10}C_3)(1/6)^3(5/6)^7$ or binompdf(10,1/6,3) = 0.155045, or about 15.5 percent.

The probability of getting less than four sixes is binomcdf(10,1/6,3) = 0.93027, or about 93 percent. Hence, the probability of getting four or more sixes in 10 rolls of a single die is about 7 percent.

If X is the number of sixes obtained when 10 dice are rolled, then

If X is the number of 6's obtained when ten dice are rolled, then $E(X) = \mu_x = 10(1/6) = 1.6667$, and
$\sigma_x = \sqrt{10(1/6)(5/6)} = 1.1785.$

Did you notice that the coin-tossing example above is also a binomial situation?
Realize that a binomial distribution can be approximated well by a normal distribution if the number of trials is sufficiently large. If n is the number of trials in a binomial setting, and if p represents the probability of "success" on each trial, then a good rule of thumb states that a normal distribution can be used to approximate the binomial distribution if np is at least 10 and n(1-p) is at

least 10.

The primary difference between a binomial random variable and a geometric random variable is what you are counting. A binomial random variable counts the number of "successes" in n trials. A geometric random variable counts the number of trials up to and including the first "success."

IV. Statistical Inference

You must be able to decide which statistical inference procedure is appropriate in a given setting. Working lots of review problems will help you.

You need to know the difference between a population parameter, a sample statistic, and the sampling distribution of a statistic.

On any hypothesis testing problem:

1. State hypotheses in words and symbols.
2. Identify the correct inference procedure and verify conditions for using it.
3. Calculate the test statistic and the *P*-value (or rejection region).
4. Draw a conclusion in context that is directly linked to your *P*-value or rejection region.

- State your hypotheses in terms of population parameters, not sample statistics.
- Use standard notation in your hypotheses: μ for population mean and $p, \pi, or\, \theta$ for population proportion.
- Don't reverse the null and alternate hypotheses. Remember, the null hypothesis is basically a statement of no effect or no difference. If you hope to show that there is a difference between two population means, then the null hypothesis should be that the population means are equal.
- It is not enough to state the conditions for the chosen inference procedure. You must show that the conditions are satisfied.

On any confidence interval problem:

1. Identify the population of interest and the parameter you want to draw conclusions about.
2. Choose the appropriate inference procedure and verify conditions for its use.
3. Carry out the inference procedure.
4. Interpret your results in the context of the problem.

You need to know the specific conditions required for the validity of each statistical inference procedure -- confidence intervals and significance tests.

Be familiar with the concepts of Type I error, Type II error, and Power of a test.

Type I error: Rejecting a null hypothesis when it is true.
P(Type I error) = α = significance level of the test

Type II error: Accepting a null hypothesis when it is false.

Power of a test: Probability of correctly rejecting a null hypothesis
Power = 1 - P(Type II error).
You can increase the power of a test by increasing the sample size or increasing the significance level (the probability of a Type I error).