

**AP STATISTICS EXAM REVIEW**  
(Based on the topic outline of the College Board)

I. Exploring Data

Categorical Data – nominal scale, names

e.g. male/female or eye color or breeds of dogs

Quantitative Data – rational scale (can +, –, ·, ÷ with numbers describing data)

e.g. weights of hamsters or amounts of chemicals in beverages

A. Graphing one variable (univariate) data —ALWAYS PLOT DATA

1. Categorical data

*Bar graphs* (bars do not touch)

*Pie charts* (percentages must sum to 100%)

2. Quantitative data – label carefully

*Dot plots* – can resemble probability curves

*Stem (& leaf) plots* – remember to put in the key (e.g. 8|2 means 82 mg. of salt)

Split stems if too many data points

Back-to-back for comparison of two samples

*Histogram* – put // for breaks in axis, use no fewer than 5 classes (bars), check to see if scale is misleading, look for symmetry & skewness

*Ogive* – cumulative frequency plot

*Time plot* – used for seasonal variation where the x-axis is time

*Box plot* – modified shows outliers

Side-by-side are good for comparing quartiles, medians and spread

B. Summary statistics for one variable data (use calculator with 1-variable stats)

1. Measures of central tendency (center)

mean ( $\bar{x}$ ,  $\mu$ )

median (middle)

mode (most)

2. Measures of dispersion (spread)

range (max – min)

quartile (25% =  $Q_1$ , 75% =  $Q_3$ )

interquartile range ( $Q_1 - Q_3$ )

$$\text{variance } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ or } \sigma^2 = \frac{\sum (\bar{x}_i - \mu)^2}{n}$$

standard deviation – square root of variance ( $s$ ,  $\sigma$ )

Mean, range, variance, and standard deviation are non-resistant measures (strongly influenced by outliers). Use variance and standard deviation with approximately normal distributions only.



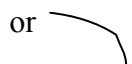
Remember: the mean chases the tail.

- C. Graphing two variable (bivariate) data – DATA MUST BE QUANTITATIVE. Graph the explanatory variable (independent) on the x axis, the response variable (dependent) on the y axis

Scatterplots look for relationships between the variables.

Look for clusters of points and gaps. Two clusters indicate that the data should be analyzed to find reasons for the clusters.

If the points are scattered, draw an ellipse around the plot. The more elongated, the stronger the linear relationship. Sketch the major axis of the ellipse. This is a good model of the linear regression line.

If the data appears curved in the shape of a power function or an exponential function,  or  or , use the calculator to fit an appropriate function to the data.

- D. Analyzing two variable quantitative data when a linear relationship is suggested

1. Linear correlation coefficient ( $r$ ) – measures the strength of the linear relationship

$$-1 \leq r \leq 1$$

$r = 0$  indicates no relationship (the ellipse is a perfect circle)

$-r$  indicates an inverse relationship

$r$  is a non-resistant measure (outliers strongly affect  $r$ )

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \quad (\text{use calculator with 2-variable stats})$$

2. Least squares regression line (LSRL) – used for prediction; minimizes the vertical distances from each data point to the line drawn. (Linreg  $a+bx$ )

$y$  varies with respect to  $x$ , so choose the explanatory and response axes carefully (y is dependent on x)

$\hat{y}$  = predicted y value

$\hat{y} = a + bx$  is the equation of the LSRL where  $b$  = slope =  $r \left( \frac{s_y}{s_x} \right)$  and the point  $(\bar{x}, \bar{y})$

is always on the line.

Do not *extrapolate* (predict a y value when the x value is far from the other x values).

3. Coefficient of Determination ( $r^2$ ) – gives the proportion (%) of variation in the values of y that can be explained by the regression line. The better the line fits, the higher the value of  $r^2$ .

To judge "fit of the line" look at  $r$  and  $r^2$ . If  $r = 0.7$ , then  $r^2 = .49$ , so about half the variation in y is accounted for by the least squares regression line.

4. Residual ( $y - \hat{y}$ ) – vertical distance from the actual data point to the regression line.

$y - \hat{y}$  = observed y value – predicted y value

residuals sum to zero

Residual plot – scatterplot of (observed x values, predicted y values) or  $(x, \hat{y})$

Use calculator to plot residuals on y axis, original x values on x axis. Check:

no pattern → good linear relationship,

curved pattern → no linear relationship,

plot widens → larger x values do not predict y values well

Outliers – y values far from the regression line (have large residuals)

Influential points – x values far from the regression line (may have small residuals)

- E. Analyzing two variable quantitative data when the data is in a curved pattern
1. Take the log (or ln) of both the original  $x$  and original  $y$  data (in your lists)
  2. Replot the data using the logarithmic transformation; it will look linear
  3. Calculate a least squares regression line as if the transformed data was original data. Remember that the regression equation is now in the form  $\log \hat{y} = a + b \log x$  (or  $\ln \hat{y} = a + b \ln x$ ).
- Remember to undo the transformation before making a prediction using the regression line.

- F. Cautions in analyzing data
1. *Correlation does not imply causation.* Only a well-designed, controlled experiment may establish causation
  2. Lurking variables (variables not identified or considered) may explain a relationship (correlation) between the explanatory and response variables by either confounding (a third variable affects the response variable only – Hawthorne effect) or by common response (a third variable affects both the explanatory and response variables – population growth affected both Methodist ministers and liquor imports).

## II. Methods of data collection

- A. Census – contacts every individual in the population to obtain data  
Symbols  $\mu$  and  $\sigma$  are *parameters* and are used only with population data

- B. Sample survey – collects data from a part of a population in order to learn about the entire population

Symbols  $\bar{x}$  and  $s_x$  are *statistics* and are used with sample data

1. Bad sampling designs result in bias in different forms
  - voluntary response sample* – participants choose themselves, usually those with strong opinions choose to respond  
e.g. on-line surveys, call-in opinion questions
  - convenience sample* – investigators choose to sample those people who are easy to reach  
e.g. marketing surveys done in a mall
  - bias* – the design systematically favors certain outcomes or responses  
e.g. surveying pacifist church members about attitudes toward war
2. Good sampling designs
  - simple random sample* – a group of  $n$  individuals chosen from a population in such a way that every set of  $n$  individuals has an equal chance of being the sample actually chosen; use a random number table or randint on the calculator
  - stratified random sample* – divide the population into groups (strata) of similar individuals (by some chosen category) then choose a simple random sample from each of the groups
  - multistage sampling design* – combines stratified and random sampling in stages
  - systematic random sampling* – choosing every  $n^{\text{th}}$  individual after choosing the first randomly

Cautions (even when the design is good) include:

  - undercoverage* – when some groups of the population are left out, often because a complete list of the population from which the sample was chosen was not available.  
e.g. U.S. census has a task force to get data from the homeless because the homeless do not have addresses to receive the census forms in the mail

*nonresponse* – when an individual appropriately chosen for the sample cannot or does not respond  
*response bias* – when an individual does not answer a question truthfully, e.g. a question about previous drug use may not be answered accurately  
*wording of questions* – questions are worded to elicit a particular response, e.g. One of the Ten Commandments states, "Thou shalt not kill." Do you favor the death penalty?

- C. Observational study – observes individuals in a population or sample, measures variables of interest, but does not in any way assign treatments or influence responses
- D. Experiment – deliberately imposes some treatment on individuals (experimental units or subjects) in order to observe response. *Can give evidence for causation if well designed with a control group.* 3 necessities: *Control – Randomize – Replicate*  
*Control* – for lurking variables by assigning units to groups that do not get the treatment  
*Randomize* – use simple random sampling to assign units to treatments/control groups  
*Replicate* – use the same treatment on many units to reduce the variation due to chance  
 The "best" experiments are double blind – neither the investigators nor the subjects know which treatments are being used on which subjects. Placebos are often used.  
 Block designs – subjects are grouped before the experiment based on a particular characteristic or set of characteristics, then simple random samples are taken within each block. (remember the tree problem?) Matched pairs is one type of block design where two treatments are assigned, sometimes to the same subject, sometimes to two different subjects matched very closely.  
 Remember how to sketch a design (p. 302)

### III. "Simple" Probability – probability only refers to "the long run", never short term

#### A. Basic definitions

1. *Independent* – one event does not change (have an effect on) another event
2. *Mutually exclusive (disjoint)*– events cannot occur at the same time, so there can be no intersection of events in a Venn diagram. Mutually exclusive events ALWAYS have an effect on each other so can never be independent.

#### B. General rules

1. All probabilities for one event must sum to 1
2.  $P(A^c) = 1 - P(A)$   $A^c$  is the complement of A
3.  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$  {or means  $\cup$  - union}
4.  $P(A \text{ and } B) = P(A) \cdot P(B|A)$  {and means  $\cap$  - intersection}
5.  $P(B|A) = P(B \text{ given } A) = \frac{P(A \text{ and } B)}{P(A)}$  (conditional probability)
6. If  $P(A \text{ and } B) = 0$  then A and B are mutually exclusive
7. If  $P(B|A) = P(B)$ , then A and B are independent
8. If  $P(A \text{ and } B) = P(A) \cdot P(B)$ , then A and B are independent

#### IV. Types of probability distributions

##### A. Probability distributions of random variables

1. Graphs, whether of continuous or discrete variables, must have area under a curve = 1. Histograms – discrete smooth curves – continuous. The graphs need not be symmetric.
2. To get the *expected value* or *mean* of a discrete random variable, multiply the number of items by the probability assigned to each item (usually given in a probability distribution table), then sum those products,  $\mu = \sum x_i p_i$
3. To get the variance of a discrete random variable, use  $\sigma^2 = \sum (x_i - \mu)^2 p_i$  where  $p$  is the probability assigned to each item,  $x$ .
4. To find the sum or difference ( $\pm$ ) using two random variables, add or subtract the means to get the mean of the sum or difference of the variables,  $\mu_{x \pm y} = \mu_x \pm \mu_y$ . To get the standard deviation of the combined variables, first determine whether or not the random variables are independent (no correlation coefficient,  $\rho$ , given.) If independent, add the *variances*, then take the square root of the sum,  $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$ . If there is a correlation given between the variables (they are not independent), add (when summing the variables) or subtract (when finding the difference between the two variables)  $2\rho\sigma_x\sigma_y$  (where  $\rho$  is the correlation coefficient) under the radical before taking the square root,  $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2 \pm 2\rho\sigma_x\sigma_y}$ .

##### B. The binomial distribution – conditions: (1) used when there are only two options (success or failure), (2) there is a fixed number of observations, (3) all observations are independent, (4) the probability of success, $p$ , is constant.

1. The mean of a binomial distribution is  $\mu = np$  where  $p$  is the probability and  $n$  is the number of observations in the sample.
2. The standard deviation of the binomial distribution is  $\sigma = \sqrt{np(1-p)}$
3. The graph of a binomial distribution is strongly right skewed (has a long right tail) unless the number in the sample is very large, then the distribution becomes normal.
4. binomial probability –  $nCr(p)^r(1-p)^{n-r}$   
e.g. the probability of choosing a certain color is .35. If 8 people are in a room, the probability that exactly 5 of those will choose the color is  $\frac{8!}{5!(8-5)!} (.35)^5 (.65)^3$ .

On calculator use 2<sup>nd</sup> DIST binompdf(8,.35, 5).

To find the probability that 5 or less will choose the color, sum the individual probabilities of 0, 1, 2, 3, 4, & 5, OR use  $1 - (\text{sum of probabilities of 6, 7, 8})$

On calculator use 2<sup>nd</sup> DIST binomcdf(8,.35,5).

##### C. The geometric distribution – conditions are the same as for the binomial except there is not a fixed number of observations because the task is to find out how many times it takes before a success occurs. This is sometimes called a waiting time distribution.

1. The mean of the geometric distribution is  $\mu = \frac{1}{p}$
2. The standard deviation of the geometric distribution is  $\sigma = \sqrt{\frac{1-p}{p^2}}$
3. The graph of the geometric distribution is strongly right skewed always.
4. geometric probability –  $(1-p)^{n-1} p$   
 e.g. the probability, when rolling a fair die, of rolling a particular number (say a 4) for the first time on the 7th roll is  $(1 - \frac{1}{6})^{7-1} (\frac{1}{6})$ . On calculator, use 2<sup>nd</sup> DIST geometpdf( $\frac{1}{6}$ , 7).  
 To find the probability that rolling a particular number will take more than 7 rolls of the die, use 1 – (the sum the geometric probabilities from 1 up to 7)  
 On calculator use  $1 - 2^{\text{nd}} \text{ DIST geometcdf}(\frac{1}{6}, 7)$ .

D. The normal distribution – the bell curve

1. If the mean = 0 and the standard deviation = 1, this is a standard, normal curve
2. Use with z scores (standard scores),  $z = \frac{\bar{x} - \mu}{\sigma}$ , where +z are scores are above the mean and –z are scores below the mean.
3. To compare two observations from different circumstances, find the z score of each, then compare
3. Use z scores to find the p value, the probability (or proportion or percent) of the data that lies under a portion of the bell curve, p values represent area under the curve. Use shadenorm and normalcdf (to find the p value), or invnorm (to find z score) on the calculator
4. ALWAYS DRAW THE CURVE and shade to show your area.
5. 68% – 95% – 99.7% rule for area under the curve

V. Sampling distributions – probability distributions which are made from the measured characteristic of several different samples, e.g. the sampling distribution of the *means*. Remember: larger sample sizes yield less variability (spread).

Look for bias and/or variability. Bias is when the sample mean is not close to the population mean, variability is when the sample means are widely scattered. Remember the targets on page 500.

A. Sampling distribution of the means – when working with the  $\bar{x}$  's of samples. Check for large large sample size,  $n(10) < \text{pop. size}$ , and a normal distribution.

1. The mean of the sampling distribution of the means equals the mean of the population.
2. The standard deviation of the sampling distribution equals  $\frac{\sigma}{\sqrt{n}}$ , also called the standard error of the mean.
3. Central limit theorem – when the number, n, in the sample size is large, the graph of the sampling distribution of the sample means looks like the normal bell curve. As n gets larger, the spread (variance and standard deviation) decreases (remember example with dates of pennies.)
4. Use z scores to find probabilities, regardless of sample size, if the question asks about the mean of a sample (not about one observation in a sample).

- B. Sampling distribution of a proportion – when working with a percentage value or proportion e.g. 38% of the population thinks frogs are icky. What is the probability that, in a sample of 120 people, only 29% thinks the same?

Check for large sample size ( $n > 30$ ) and that  $np > 10$  and  $n(1-p) > 10$  in order to use a normal distribution.

1. The mean of the sampling distribution equals the mean of the population,  $\mu_{\hat{p}} = p$ .
2. The standard deviation of the sampling distribution  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ .
3. Use z scores to find probabilities because probabilities (proportions/percentages) are used

VI. Confidence intervals – "I am \_\_\_% confident that the population mean lies between the upper and lower levels of the confidence interval" Remember it either does, or does not! Review the picture on page 541 – if many samples are taken, the means of \_\_\_% of them should lie within the confidence interval. Confidence level means \_\_%, confidence interval gives the lower and upper upper bounds.

- A. State that the data came from a simple random sample from the population of interest and that the sampling distribution is approximately normal. If the size of the sample is small, show that the sample is normal by using a stem and leaf plot, a histogram, or the standard normal probability plot on the calculator (with the data in a list). Outliers can strongly affect this – delete an outlier from the data to compute confidence (but always state that you did so and why).

B. Confidence intervals are *always* two-tailed.

C. The confidence interval for the population mean  $\mu = \bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$  or  $\mu = \hat{p} \pm z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

D. The confidence interval for the sample mean  $\bar{x} = \mu \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$  or  $\hat{p} = p \pm z\left(\sqrt{\frac{p(1-p)}{n}}\right)$

E. Margin of error =  $\pm z\left(\frac{\sigma}{\sqrt{n}}\right)$ ; the z score corresponding to a confidence level where  $\frac{\alpha}{2} = \frac{1-\text{___}\%}{2}$

1. Use the margin of error formula to solve for n, the number needed in the sample to yield a particular level of confidence.
2. To decrease the margin of error, increase the sample size, decrease the standard deviation, or decrease the confidence interval.

VII. Significance tests – require a null hypothesis containing  $\mu$  and = and an alternative hypothesis

containing  $\mu$  and either  $\neq$ ,  $<$ , or  $>$ ; e.g.  $H_0: \mu = 0$ ,  $H_a: \mu < 0$ . Some hypotheses are words only.

Determine whether the test is two tailed ( $H_a: \mu \neq 0$ ) or one tailed ( $H_a: \mu < 0$  or  $H_a: \mu > 0$ ) Choose a rejection  $\alpha$  level, e.g.  $\alpha = .05$ . Reject the null hypothesis if the test statistic falls in the tail or if the p value is smaller than a chosen  $\alpha$  level. Fail to reject  $H_0$  if the test statistic is not in the tail or if the p value is larger than a chosen  $\alpha$  level. Choose the test carefully and always state the conditions of the test before beginning to test. Use your TEST BIBLE! Memorize conditions for each test! Always refer to the problem in context when addressing conditions, such as simple random sample. Write the conclusion based on your test – reject or fail to reject the null hypothesis using the problem situation.

- A. z test – use with means,  $\bar{x}$  and  $\mu$ , when the population standard deviation,  $\sigma$ , is given
1. Conditions – simple random sample from population given, normal population or large sample size, If the sample is large, then it can be assumed to be normal. If not, show normality with a stem and leaf plot, a histogram, or the standard normal probability plot on the calculator (with the data in a list).
  2.  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ . Use calculator Z-test.
  3. Write conclusion in terms of the problem.
- B. Simple t test – use with means,  $\bar{x}$  and  $\mu$ , when  $\sigma$  is not known
1. Conditions – If the sample size is very small (<15) use this test only when the sample is normal (be sure to show normality). If the sample size is between 15 and about 30, omit outliers, show normality, and use this test. If the sample size is >30, simply state that the test can be used because of a large sample size.
  2. Use  $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$  Remember to state degrees of freedom =  $n - 1$ . Use calculator T-Test.
  3. Write conclusion in terms of the problem.
- C. *Dependent* or *matched pairs* t test
1. Conditions – same as t above but the same subjects in the sample receive two different treatments.
  2. First find the differences (subtract) between the two treatments for each subject, then find the mean and standard deviation of the differences. The null hypothesis is always that the differences equal 0.
  3. Use  $t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}}$  with  $n - 1$  degrees of freedom to test. Work as the t test above.
- D. Two *independent* samples t test
1. Conditions – same as simple t but two different samples are compared. The standard deviations of each sample must be reasonably similar.
  2. Find the mean and standard deviations of each of the samples.
  3.  $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  Use calculator with 2-sampTTest to get p value and degrees of freedom
  4. Write conclusion in terms of the problem.
- E. Test for proportions (percentages) - always use Z-Prop
1. Conditions – use rules of thumb with 10:  $n(10) < \text{population size}$ ,  $n(p) > 10$ ,  $n(1-p) > 10$
  2. If the population proportion is given, use  $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$  Use calculator 1-PropZTest
  3. If sample size is large,  $\hat{p}$  is approximately equal to  $p_0$ , the population proportion.



F. Test for comparing proportions (percentages) - always use Z-Prop

1. Conditions – use rules of thumb with 10 with both sample proportions. Samples must be independent.

2. Use  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$  or, if given raw numbers so that the totals for the two

samples can be added to get one  $\hat{p}$  that applies to the *pooled* sample data,

use  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ , and the standard z procedures.

G. Chi – squared goodness of fit/homogeneity of proportions/independence (association) tests – analyzing two variable categorical data with data in a matrix. Elements in each cell are frequencies – the number of individual (objects) with those particular characteristics

Definitions: marginal distribution – computing the totals for the row categories and column categories (may compute the percentage using a row (or column) total divided by the “total total”); conditional distribution – computing the percent using the number in a cell divided by the row (or column) total.

1. Conditions – simple random sample and all expected counts are  $\geq 5$ .  
Hypothesis in words: expected values are as claimed/there is no difference/there is no association
2.  $\chi^2$  can be used to determine an expected number of outcomes by using the statistic below or in a two way table by locating the cell of interest and using  $\frac{\text{rowtotal} \cdot \text{columntotal}}{\text{"total" total}}$
3. With a matrix of observed values, put matrix in your calculator. Use  $\chi^2$  Test on calculator. The expected matrix will appear in matrix B.
4.  $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$  with  $n - 1$  degrees of freedom if only one row in the matrix or  $(r - 1)(c - 1)$  degrees of freedom if the matrix has more than one row (calculator will state this)

H. Linear Regression Tests – use only after computing a regression line

1. Conditions – y responses are independent, the relationship between x and y is linear, standard deviations about the regression line are constant, there is normal variation (normal curve about each predicted y value)
2. Use LinRegTTest on calculator
3. Use standard error formula on ap formula sheets

I. Definitions in significance testing

1. Type I error =  $\alpha$  level, the probability the null hypothesis will be rejected when it is true
2. Type II error =  $\beta$  level, the probability the null hypothesis will be accepted when it is false.
3. The power of any test is  $1 - \beta$ , the probability of rejecting the null hypothesis when it is false (making the correct decision to reject). The power increases when the  $\alpha$  level is larger, therefore a test with  $\alpha = .05$  is more powerful than a test with  $\alpha = .01$  because an investigator is more likely to reject the null hypothesis. That means that the power increases when the confidence interval gets smaller, since a larger  $\alpha$  level means a smaller confidence interval. The power also increases when the sample size is increased because the sample more accurately reflects the population.
4. A test is robust when it answers the question of reject/not reject correctly even if all of the assumptions of the test are not met. e.g. a t-test which rejects the null hypothesis correctly even with a small sample size.

EXAM TAKING TIPS:

<http://apcentral.collegeboard.com/members/article/1,3046,152-172-0-4073,00.html>