AP Statistics
Cumulative AP Exam Study Guide

Statistics – the science of collecting, analyzing, and drawing conclusions from data.
- Descriptive – methods of organizing and summarizing statistics
- Inferential – making generalizations from a sample to the population.

Population – an entire collection of individuals or objects.

Sample – A subset of the population selected for study.

Variable – any characteristic whose value changes.

Data – observations on single or multi-variables.

Variables
- Categorical – (Qualitative) – basic characteristics
- Numerical – (Quantitative) – measurements or observations of numerical data.
  - Discrete – listable sets (counts)
  - Continuous – any value over an interval of values (measurements)
- Univariate – one variable
- Bivariate – two variables
- Multivariate – many variables

Distributions
- Symmetrical – data on which both sides are fairly the same shape and size. "Bell Curve"
- Uniform – every class has an equal frequency (number) "a rectangle"
- Skewed – one side (tail) is longer than the other side. The skewness is in the direction that the tail points (left or right)
- Bimodal – data of two or more classes have large frequencies separated by another class between them. "double hump camel"

How to describe numerical graphs - S.O.C.S.
- Shape – overall type (symmetrical, skewed right left, uniform, or bimodal)
- Outliers – gaps, clusters, etc.
- Center – middle of the data (mean, median, and mode)
- Spread – refers to variability (range, standard deviation, and IQR)
- *Everything must be in **context** to the data and situation of the graph.
- *When comparing two distributions – MUST use comparative language!

Parameter – value of a population (typically unknown)

Statistic – a calculated value about a population from a sample(s).

Measures of Center
- Median – the middle point of the data (50th percentile) when the data is in numerical order. If two values are present, then average them together.
- Mean – μ is for a population (parameter) and x is for a sample (statistic).
- Mode – occurs the most in the data. There can be more then one mode, or no mode at all if all data points occur once.

Variability – allows statisticians to distinguish between usual and unusual occurrences.

Measures of Spread (variability)

Range – a single value – (Max – Min)

IQR – interquartile range – (Q3 – Q1)

Standard deviation – σ for population (parameter) & s for sample (statistic) – measures the typical or average deviation of observations from the mean – sample standard deviation is divided by df = n-1

*Sum of the deviations from the mean is always zero!

Variance – standard deviation squared

Resistant – not affected by outliers.

Resistant	Non-Resistant
Median	Mean
IQR	Range
	Variance
	Standard Deviation
	Correlation Coefficient (r)
	Least Squares Regression Line (LSRL)
	Coefficient of Determination $\left(r^2\right)$

Comparison of mean & median based on graph type

Symmetrical – mean and the median are the same value.

Skewed Right – mean is a larger value than the median.

Skewed Left – the mean is smaller than the median.

*The mean is always pulled in the direction of the skew away from the median.

Trimmed Mean – use a % to take observations away from the top **and** bottom of the ordered data. This possibly eliminates outliers.

Linear Transformations of random variables

$\mu_{a+bx} = a + b\mu_x$ The mean is changed by **both** addition (subtract) & multiplication (division).

$\sigma_{a+bx} = \sigma_{a+bx} = |b|x$ The standard deviation is changed by multiplication (division) ONLY.

Combination of two (or more) random variables

$\mu_{x\pm y} = \mu_x \pm \mu_y$ Just add or subtract the two (or more) means

$\sigma_{x\pm y} = \sqrt{\sigma_x^2 + \sigma_y^2}$ Always add the variances – X & Y MUST be independent

Z –Score – is a standardized score. This tells you how many standard deviations from the mean an observation is. It creates a standard normal curve consisting of z-scores with a $\mu = 0$ & $\sigma = 1$.

$$\mathbf{z} = \frac{\mathbf{x} - \mu}{\sigma}$$

Normal Curve – is a bell shaped and symmetrical curve.

As σ increases the curve flattens.

As σ decreases the curve thins.

Empirical Rule (68-95-99.7) measures 1σ, 2σ, and 3σ on **normal curves** from a center of μ.

68% of the population is between -1σ and 1σ

95% of the population is between -2σ and 2σ

99.7% of the population is between -3σ and 3σ

Boxplots – are for medium or large numerical data. It does not contain original observations. Always use modified boxplots where the fences are 1.5 IQRs from the ends of the box (Q1 & Q3). Points outside the fence are considered outliers. Whiskers extend to the smallest & largest observations within the fences.

5-Number Summary – Minimum, Q1 (1st Quartile – 25th Percentile), Median, Q3 (3rd Quartile – 75th Percentile), Maximum

Probability Rules

Sample Space – is collection of **all** outcomes.

Event – any sample of outcomes.

Complement – all outcomes **not** in the event.

Union – A or B, all the outcomes in both circles. $A \cup B$

Intersection – A and B, happening in the middle of A and B. $A \cap B$

Mutually Exclusive (Disjoint) – A and B have no intersection. They cannot happen at the same time.

Independent – if knowing one event does not change the outcome of another.

Experimental Probability – is the number of success from an experiment divided by the total amount from the experiment.

Law of Large Numbers – as an experiment is repeated the experimental probability gets closer and closer to the true (theoretical) probability. The difference between the two probabilities will approach "0".

Rules

(1) All values are $0 < P < 1$.

(2) Probability of sample space is 1.

(3) Compliment = P + (1 - P) = 1

(4) Addition P(A or B) = P(A) + P(B) – P(A & B)

(5) Multiplication P(A & B) = P(A) · P(B) if a & B are independent

(6) P (at least 1 or more) = 1 – P (none)

(7) Conditional Probability – takes into account a certain condition. $P(A|B) = \dfrac{P(A \& B)}{P(B)} = \dfrac{P(both)}{P(given)}$

Correlation Coefficient – (r) – is a quantitative assessment of the strength and direction of a linear relationship. (use ρ (rho) for population parameter)

Values – [-1, 1] 0 – no correlation, (0, ±0.5) – weak, [±0.5, ±0.8) – moderate, [±0.8, ±1] - strong

Least Squares Regression Line (LSRL) – is a line of mathematical best fit. Minimizes the deviations (residuals) from the line. Used with bivariate data.

$\hat{y} = a + bx$ x is independent, the explanatory variable & y is dependent, the response variable

Residuals (error) – is vertical difference of a point from the LSRL. All residuals sum up to "0".

Residual = $y - \hat{y}$

Residual Plot – a scatterplot of (x (or $\hat{y}$) , residual). No pattern indicates a linear relationship.

Coefficient of Determination $\left(r^2\right)$ - gives the proportion of variation in y (response) that is explained by the relationship of (x, y). Never use the adjusted r^2.

Interpretations: must be in context!

Slope (b) – For unit increase in x, then the y variable will increase/decrease slope amount.

Correlation coefficient (r) – There is a strength, direction, linear association between x & y.

Coefficient of determination $\left(r^2\right)$ - Approximately r^2% of the variation in y can be explained by the LSRL of x any y.

Extrapolation – LRSL cannot be used to find values outside of the range of the original data.

Influential Points – are points that if removed significantly change the LSRL.

Outliers – are points with large residuals.

Census – a complete count of the population. Why not to use a census?
- Expensive
- Impossible to do
- If destructive sampling you get extinction

Sampling Frame – is a list of everyone in the population.

Sampling Design – refers to the method used to choose a sample.

SRS (Simple Random Sample) – one chooses so that each unit has an equal chance **and** every set of units has an equal chance of being selected.
Advantages: easy and unbiased
Disadvantages: large $\sigma 2$ and must know population.

Stratified – divide the population into homogeneous groups called strata, then SRS each strata.
Advantages: more precise than an SRS and cost reduced if strata already available.
Disadvantages: difficult to divide into groups, more complex formulas & must know population.

Systematic – use a systematic approach (every 50th) after choosing randomly where to begin.
Advantages: unbiased, the sample is evenly distributed across population & don't need to know population.
Disadvantages: a large $\sigma 2$ and can be confounded by trends.

Cluster Sample – based on location. Select a random location and sample ALL at that location.
Advantages: cost is reduced, is unbiased& don't need to know population.
Disadvantages: May not be representative of population and has complex formulas.

Random Digit Table – each entry is equally likely and each digit is independent of the rest.

Random # Generator – Calculator or computer program

Bias – Error – favors a certain outcome, has to do with center of sampling distributions – if centered over true parameter then considered unbiased

Sources of Bias
- Voluntary Response – people choose themselves to participate.
- Convenience Sampling – ask people who are easy, friendly, or comfortable asking.
- Undercoverage – some group(s) are left out of the selection process.
- Non-response – someone cannot or does not want to be contacted or participate.
- Response – false answers – can be caused by a variety of things
- Wording of the Questions – leading questions.

Experimental Design

Observational Study – observe outcomes with out giving a treatment.
Experiment – actively imposes a treatment on the subjects.
Experimental Unit – single individual or object that receives a treatment.
Factor – is the explanatory variable, what is being tested
Level – a specific value for the factor.
Response Variable – what you are measuring with the experiment.
Treatment – experimental condition applied to each unit.
Control Group – a group used to compare the factor to for effectiveness – does NOT have to be placebo
Placebo – a treatment with no active ingredients (provides control).
Blinding – a method used so that the subjects are unaware of the treatment (who gets a placebo or the real treatment).
Double Blinding – neither the subjects nor the evaluators know which treatment is being given.

Principles

Control – keep all extraneous variables (not being tested) constant
Replication – uses many subjects to quantify the natural variation in the response.
Randomization – uses chance to assign the subjects to the treatments.

The only way to show cause and effect is with a **well designed, well controlled** experiment.

Experimental Designs

Completely Randomized – all units are allocated to all of the treatments randomly

Randomized Block – units are blocked and then randomly assigned in each block –reduces variation

Matched Pairs – are matched up units by characteristics and then randomly assigned. Once a pair receives a certain treatment, then the other pair automatically receives the second treatment. **OR** individuals do both treatments in random order (before/after or pretest/post-test). Assignment is dependent

Confounding Variables – are where the effect of the variable on the response cannot be separated from the effects of the factor being tested – happens in observational studies – when you use random assignment to treatments you do NOT have confounding variables!

Randomization – reduces bias by spreading extraneous variables to all groups in the experiment.

Blocking – helps reduce variability. Another way to reduce variability is to increase sample size.

Random Variable – a numerical value that depends on the outcome of an experiment.

Discrete – a count of a random variable

Continuous – a measure of a random variable

Discrete Probability Distributions -gives values & probabilities associated with each possible x.

$\mu_X = \sum x_i p(x_i)$ and $\sigma_x = \sqrt{\sum (x_i - \mu_X)^2 p(x_i)}$ calculator shortcut – 1 VARSTAT L1,L2

Fair Game – a fair game is one in which all pay-ins equal all pay-outs.

Special discrete distributions:

Binomial Distributions

Properties- two mutually exclusive outcomes, fixed number of trials (n), each trial is independent, the probability (p) of success is the same for all trials,

Random variable - is the number of successes out of a fixed # of trials. Starts at X = 0 and is finite.

$\mu_X = np \quad \sigma_x = \sqrt{npq}$

Calculator: binomialpdf (n, p, x) = single outcome $P(X = x)$

binomialcdf (n, p, x) = cumulative outcome $P(X < x)$

1 - binomialcdf (n, p, (x -1)) = cumulative outcome $P(X > x)$

Geometric Distributions

Properties -two mutually exclusive outcomes, each trial is independent, probability (p) of success is the same for all trials. (NOT a fixed number of trials)

Random Variable –when the FIRST success occurs. Starts at 1 and is ∞.

Calculator: geometricpdf (*p*, *a*) = single outcome $P(X = a)$

geometriccdf (*p*, *a*) = cumulative outcomes $P(X < a)$

1 - geometriccdf (*n*, *p*, (*a* -1)) = cumulative outcome $P(X > a)$

Continuous Random Variable -numerical values that fall within a range or interval (measurements), use density curves where the area under the curve always = 1. To find probabilities, find area under the curve

Unusual Density Curves -any shape (triangles, etc.)

Uniform Distributions –uniformly (evenly) distributed, shape of a rectangle

Normal Distributions -symmetrical, unimodal, bell shaped curves defined by the parameters μ & σ

Calculator: Normalpdf – used for graphing only

Normalcdf(lower bound, upper bound, μ, σ) – finds probability

InvNorm(p) – z-score OR InvNorm(p, μ, σ) – gives x-value

To assess Normality - Use graphs – dotplots, boxplots, histograms, or normal probability plot.

Distribution – is all of the values of a random variable.

Sampling Distribution – of a statistic is the distribution of all possible values of all possible samples. Use normalcdf to calculate probabilities – be sure to use correct SD

$\mu_{\bar{X}} = \mu_X$ $\qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ (standard deviation of the sample means)

$\mu_{\hat{p}} = p$ $\qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$ (standard deviation of the sample proportions)

$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2}$ $\qquad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ (standard deviation of the difference in sample means)

$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ $\qquad \sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$ (standard deviation of the difference in sample proportions)

$\mu_b = \beta$ $\qquad s_{b_1}$ (do not need to find, usually given in computer printout (standard error of the slopes of the LSRLs)

Standard error – estimate of the standard deviation of the statistic

Central Limit Theorem – when n is sufficiently large (n > 30) the sampling distribution is approximately normal even if the population distribution is not normal.

Confidence Intervals

Point Estimate – uses a single statistic based on sample data, this is the simplest approach.

Confidence Intervals – used to estimate the unknown population parameter.

Margin of Error – the smaller the margin of error, the more precise our estimate

Steps:

Assumptions – see table below

Calculations – C.I. = statistic ± critical value (standard deviation of the statistic)

Conclusion – Write your statement in context.

We are [x]% confident that the true [parameter] of [context] is between [a] and [b].

What makes the margin of error smaller

- make critical value smaller (lower confidence level).
- get a sample with a smaller s.
- make n larger.

T distributions compared to standard normal curve

- centered around 0
- more spread out and shorter
- more area under the tails.
- when you increase n, t-curves become more normal.
- can be no outliers in the sample data
- Degrees of Freedom = n – 1

Robust – if the assumption of normality is not met, the confidence level or p-value does not change much – this is true of t-distributions because there is more area in the tails

Hypothesis Tests

Hypothesis Testing – tells us if a value occurs by random chance or not. If it is unlikely to occur by random chance then it is statistically significant.

Null Hypothesis – H_0 is the statement being tested. Null hypothesis should be "no effect", "no difference", or "no relationship"

Alternate Hypothesis – Ha is the statement suspected of being true.

P-Value – assuming the null is true, the probability of obtaining the observed result or more extreme

Level of Significance – α is the amount of evidence necessary before rejecting the null hypothesis.

Steps:

- Assumptions – see table below
- Hypotheses - don't forget to define parameter
- Calculations – find z or t test statistic & p-value
- Conclusion – Write your statement in context.
  - Since the p-value is < (>) α, I reject (fail to reject) the Ho. There is (is not) sufficient evidence to suggest that [Ha].

Type I and II Errors and Power

Type I Error – is when one rejects H0 when H0 is actually true. (probability is α)
Type II Error – is when you fail to reject H0, and H0 is actually false. (probability is β)

α and β are inversely related. Consequences are the results of making a Type I or Type II error. Every decision has the possibility of making an error.

The Power of a Test – is the probability that the test will reject the null hypothesis when the null hypothesis is false assuming the null is true. Power = 1 – β

If you increase	Type I error	Type II error	Power
α	Increases	Decreases	Increases
n	Same	Decreases	Increases
$(\mu_0 - \mu_a)$	Same	Decreases	Increases

χ2 Test – is used to test counts of categorical data.

Types

- -Goodness of Fit (univariate)
- -Independence (bivariate)
- -Homogeneity (univariate 2 (or more) samples)

χ2 distribution – All curves are skewed right, every df has a different curve, and as the degrees of freedom increase the χ2 curve becomes more normal.

Goodness of Fit – is for univariate categorical data from a single sample. Does the observed count "fit" what we expect. Must use list to perform, df = number of the categories – 1, use χ2cdf (χ2, ∞, df) to calculate p-value

Independence – bivariate categorical data from one sample. Are the two variables independent or dependent? Use matrices to calculate

Homogeneity -single categorical variable from 2 (or more) samples. Are distributions homogeneous? Use matrices to calculate

For both χ2 tests of independence & homogeneity:

$$\text{Expected counts} = \frac{(row\ total)(column\ total)}{grand\ total} \quad \& \ df = (r-1)(c-1)$$

Regression Model:

- X & Y have a linear relationship where the true LSRL is $\mu y = \alpha + \beta x$
- The responses (y) are normally distributed for a given x-value.
- The standard deviation of the responses (σy) is the same for all values of x.
    - S is the estimate for σy

Confidence Interval $b \pm t^* s_b$

Hypothesis Testing $t = \dfrac{b - \beta}{s_{b_1}}$

Assumptions:

Proportions **z - procedures**	**Means** **t - procedures**	**Counts** **$\chi 2$ - procedures**
One sample: • SRS from population • Can be approximated by normal distribution if n(p) & n(1 – p) > 10 • Population size is at least 10n	One sample: • SRS from population • Distribution is approximately normal o Given o Large sample size o Graph of data is approximately symmetrical and unimodal with no outliers	All types: • Reasonably random sample(s) • All expected counts > 5 o Must show expected counts
Two samples: • 2 independent SRS's from populations (or randomly assigned treatments) • Can be approximated by normal distribution if $n_1(p_1)$, $n_1(1 - p_1)$, n_2p_2, & $n_2(1 - p_2) > 10$ • Population sizes are at least 10n	Matched pairs: • SRS from population • Distribution of differences is approximately normal - Given - Large sample size - Graph of differences is approximately symmetrical and unimodal with no outliers	**Bivariate Data:** **t – procedures on slope** • SRS from population • There is linear relationship between x & y. • Residual plot has no pattern. • The standard deviation of the responses is constant for all values of x. • Points are scattered evenly across the LSRL in the scatterplot. • The responses are approximately normally distributed. • Graph of residuals is approximately symmetrical & unimodal with no outliers.
	Two samples: • 2 independent SRS's from populations (or randomly assigned treatments) • Distributions are approximately normal o Given o Large sample sizes o Graphs of data are approximately symmetrical and unimodal with no outliers	