AP STATISTICS
MIDTERM REVIEW

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 ($Q_1 - Q_3$)
variance $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} or \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$
standard deviation – square root of variance (s, σ)

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.

The Normal Distributions

1. If the mean = 0 and the standard deviation = 1, this is a standard, normal curve
2. Use with z scores (standard scores), z = $\frac{x-\mu}{\sigma}$, 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
4. 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
5. ALWAYS DRAW THE CURVE and shade to show your area.
6. 68% – 95% – 99.7% rule for area under the curve

Examining Relationships

A. 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

1. Scatterplots look for relationships between the variables.
2. Look for clusters of points and gaps. Two clusters indicate that the data should be analyzed to find reasons for the clusters.
3. 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.

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

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

   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)

   $r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ (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 = $r(\frac{s_y}{s_x})$ 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 where 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 $\rightarrow$ good linear relationship,
   - curved pattern $\rightarrow$ no linear relationship,
   - plot widens $\rightarrow$ larger x values do not predict y values well
   - Outliers – y values far from the regression line (have large residuals)
   - Influential points –

A. 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).

B. Relations in categorical data
   1. From a two-way table of counts, find marginal and categorical distributions (in percents)
   2. Describe relationship between two categorical variables by comparing percents
   3. Recognize Simpson's paradox and be able to explain it

Producing Data

A. Census – contacts every individual in the population to obtain data

   Symbols μ and σ 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^{th} individual after choosing the first randomly

3. 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 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.

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.

Probability: The Study of Randomness

A.

Basic definitions

1. Probability only refers to “the long run”; never short term
2. *Independent* – one event does not change (have an effect on) another event
3. *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 they 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 or B) = P(A) + P(B) – P(A and B) {or means $\cup$ - union}
   4. P(A and B) = P(A)·P(B|A) {and means $\cap$ - intersection}
   5. P(B|A) = P(B given A) = $\frac{P(AandB)}{P(A)}$ (conditional probability)
   6. If P(A 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 and B) = P(A) ·P(B), then A and B are independent

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, add the *variances*, then take the square root of the sum, $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$.

The Binomial and Geometric Distributions

A. 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 2nd 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 – (sum of probabilities of 6, 7, 8)
   On calculator use 2nd DIST binomcdf(8,.35,5).

B. 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 graph of the geometric distribution is strongly right skewed always.

3. 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 2nd 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 – 2nd DIST geometcdf($\frac{1}{6}$,7).