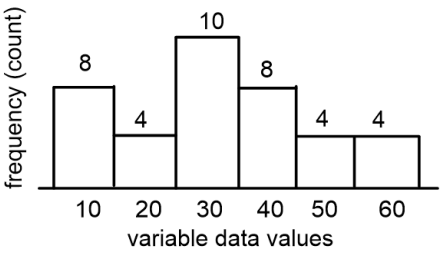
Study Guide for AP Statistics Semester 1 Final Exams

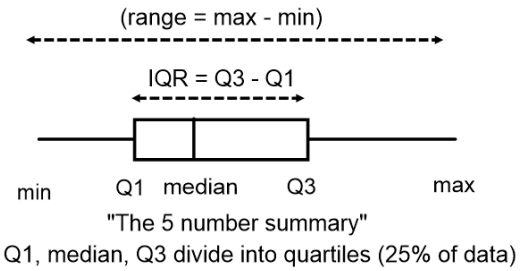
Unit 1:

Histograms



You can put data values in L1, counts in L2, 1-Var Stats to get statistics from a histogram (use value in middle of column to represent that column).

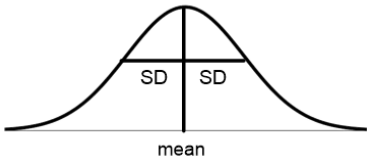
Boxplots:



Stemplots (preserves original data, can show symmetry, outliers, gaps and clusters)

2	1 1 1 3
3	2 1
4	1 0 1 6 7 3
5	
6	8

Normal distributions:



normalcdf (lower, upper, mean, SD) = area between boundaries
invNorm(leftarea, mean, SD) = upper boundary
(label these values in FRQs)

population symbols: $mean = \mu$, $standard\ deviation = \sigma$

sample statistic symbols: $mean = \overline{X}$, $standard\ deviation = s$

Standard deviation is the average distance values are from the mean.

Describing distributions: SOCS (Shape, Outliers, Center, Spread)

If distribution is skewed or has outliers you must use median for center and IQR for spread ('resistant statistics')
If distribution is symmetrical you can choose either mean or median for center, SD or IQR for spread.

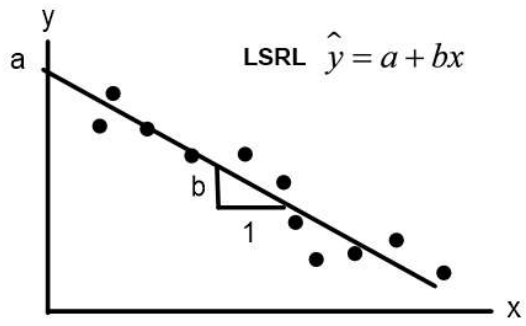
Low outlier if $< Q1 - 1.5(IQR)$, High outlier if $> Q3 + 1.5(IQR)$

Z-score = number of standard deviations above (+) or below (-) a value is from the mean. $Z = \frac{x - \mu}{\sigma}$

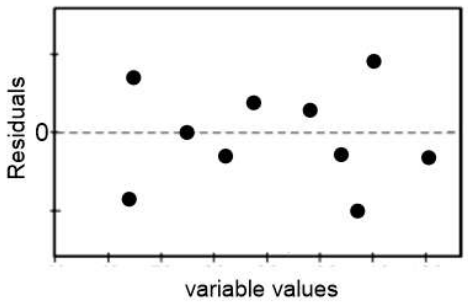
(If you use z-score in a normal distribution, mean = 0, SD = 1)

Unit 2:

scatterplot:



residual plot (magnifies differences around the LSRL)



A good residuals plot has no pattern in the residuals (linear model is appropriate).

$residual = y - \hat{y}$

Correlation Coefficient:

$r = \frac{1}{n-1}\sum Z_x Z_y$ is a measure of how far points are (in y direction) from the LSRL $\quad -1 \le r \le 1$

r^2 = coefficient of determination = % of the variation in y that is explained by the LSRL which relates y to x.

Slope for LSRL: $b = r\frac{s_y}{s_x}$ For an outlier to affect slope of LSRL it must have leverage - horiz distance from $(\bar{x}, \bar{y})$.

Unit 3:

Sampling Techniques:
- Simple Random Sample (SRS): Select a random subset from the entire population all at once (equally likely).
- Cluster Sampling: Randomly select one or more groups (clusters) and use all in each cluster for the sample.
- Stratified Random Sampling: Select an SRS from each group (strata) to include in the sample.
- Systematic Sampling: Employ a procedure (e.g. every other one) instead of randomness.

Biases:
- Undercoverage Bias: A portion of the population could not be included in the sample.
- Voluntary Response Bias: No preselection of sample – ask for volunteers to join the sample.
- Nonresponse Bias: Researchers choose a sample, but people can opt-out.
- Response Bias: Something about the survey or the way it is used makes people change their responses.

Experiments require treatment imposed by researchers on groups (control of a factor)
Good experiments also include: random assignment to groups, control of a factor (by applying a treatment which is different between at least two groups), replication (numbers in the groups).
Optional: You can block on known differences in the subjects (but not required) – reduces variation in response variable. You can employ blinding (subject and/or researchers don't know which subjects receive which treatments) which may require a placebo (fake treatment).

Explanatory variables are called factors (which each have a number of levels). Experimental units/subjects are what receive the treatments and produce results. What is measured is called the Response Variable.

Studies which don't meet criteria for experiment are called observational studies.

Unit 4:

Conditional Probability:

$$P(video\ games \mid girl) = \frac{12}{60} = .20$$

event: The event is always contained within the conditional sample space.

condition: The condition is always just a portion of the sample space (the *conditional sample space*).

The event goes in the numerator of the fraction.

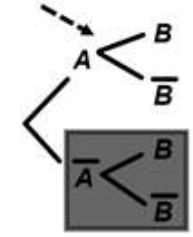
$$P(video\ games \mid girl) = \frac{12}{60} = .20$$

The condition goes in the denominator of the fraction.

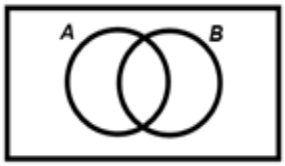
The **conditional sample space** is a portion of the **sample space**

The **event** is a portion of the **conditional sample space**.

	girls	boys	
read a book	18	4	22
video games	12	20	32
watch Netflix	30	16	46
	60	40	100



OR is add (but must subtract any overlap):



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

Special Case for OR:
Disjoint (Mutually-Exclusive) Events

$$P(A \cap B) = 0$$

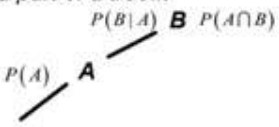
So the OR formula is simplified...

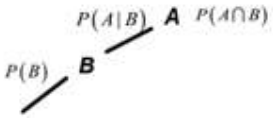
$$P(A \cup B) = P(A) + P(B)$$

AND is multiply (but 2nd probability must be conditional):

$$P(A \cap B) = P(A) \cdot P(B \mid A) \qquad P(A \cap B) = P(B) \cdot P(A \mid B)$$

Picture a part of a tree...





Special Case for AND: Independent Events

Test for independent events:

Two events are independent if:

$$P(B) = P(B \mid A) = P(B \mid \overline{A})$$

(check any two)

For independent events: $P(B) = P(B \mid A)$

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$ ...simplifies to...

$$P(A \cap B) = P(A) \cdot P(B)$$

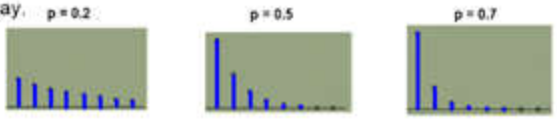
8: Discrete Probability Models

Binomial Shape depends upon p.

$\mu = np \quad \sigma = \sqrt{npq}$

Geometric Shape is always exponential decay.

$\mu = \frac{1}{p} \quad \sigma = \frac{\sqrt{1-p}}{p}$



General Discrete Models

$\mu = \text{'expected value'} = \sum X \cdot P(X)$

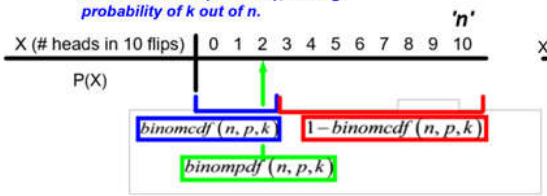
$\sigma \text{ (and } \mu\text{) found using } L1(data), L2(freqList), 1\text{-}Var\ Stats$

8: Discrete Probability Models

Binomial

- Only 2 outcomes
- Probabilities must be the same each trial.
- Probabilities of trials must be independent.
- Must have fixed number of trials, n

***Best for:** independent trials, fixed number of trials (known n), finding probability of k out of n.*

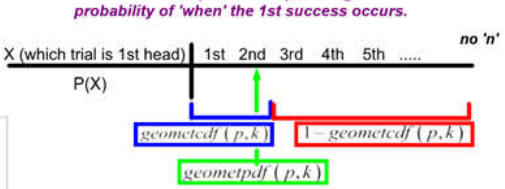


$P(\text{exactly } k \text{ successes out of } n \text{ trials}) = {}_nC_k (p)^k (q)^{n-k}$

Geometric

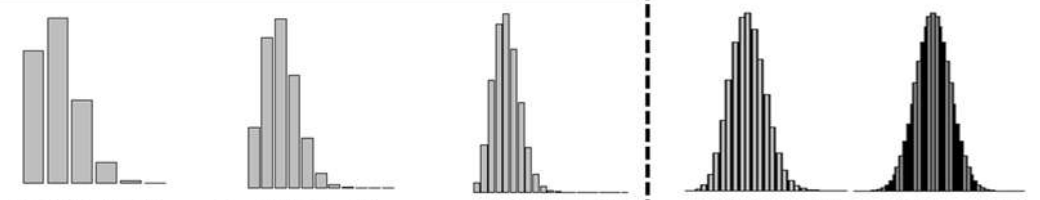
- Only 2 outcomes
- Probabilities must be the same each trial.
- Probabilities of trials must be independent.
- May or may not have fixed number of trials, n

***Best for:** independent trials, non-fixed number of trials (unknown n), finding probability of 'when' the 1st success occurs.*



$P(\text{success on the } k^{th} \text{ trial}) = (q)^{k-1} (p)$

10: Normal Approximation of Binomial Model



$p = 0.2, n = 5$ $(np = 1)$

$p = 0.2, n = 10$ $(np = 2)$

$p = 0.2, n = 20$ $(np = 4)$

$p = 0.2, n = 50$ $(np = 10)$

$p = 0.2, n = 100$ $(np = 20)$

Can use Normal approximation for the Binomial distribution.

If $np \geq 10$ *and* $nq \geq 10$

a Binomial distribution can be approximated with a Normal distribution with: $\mu = np$

$\sigma = \sqrt{npq}$

11: Combining Multiple Distributions

Define an algebraic expression for how the source distributions are used to build the new distribution:

$E = A + B - C - D$

The means are always determined by the defining algebraic expression:

$\mu_E = \mu_A + \mu_B - \mu_C - \mu_D$

But because each source of variability increases overall variation, the variances always add:

$\sigma_E^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2$

However, we must know for certain that the variables are all varying independently of one another. (If not independent, we can find mean but not standard deviation).

11: Transforming a Single Distribution

Multiplying/dividing affects both center and spread...

Adding/Subtracting affects only center...

If $Y = aX \pm b$ $\quad \mu_Y = a\mu_X \pm b$ $\quad \sigma_Y = a\sigma_X$