Statistics Cheat Sheet

Mr. Roth , Mar 2004

1. Fundamentals

a. Population – Everybody to be analysed
   - ✧ Parameter - # summarizing Pop

b. Sample – Subset of Pop we collect data on
   - ✧ Statistics - # summarizing Sample

c. Quantitative Variables – a number
   - ✧ Discrete – countable (# cars in family)
   - ✧ Continuous – Measurements – always # between

d. Qualitative
   - ✧ Nominal – just a name
   - ✧ Ordinal – Order matters (low, mid, high)

Choosing a Sample

- Sample Frame – list of pop we choose sample from
- Biased – sampling differs from pop characteristics.
- Volunteer Sample – any of below three types may end up as volunteer if people choose to respond.

<u>Sample Designs</u>

e. Judgement Samp: Choose what we think represents
   - ✧ Convenience Sample – easily accessed people

f. Probability Samp: Elements selected by Prob
   - ✧ Simple random sample – every element = chance
   - ✧ Systematic sample – almost random but we choose by method

g. Census – data on every everyone/thing in pop

<u>Stratified Sampling</u>

Divide pop into subpop based upon characteristics

h. Proportional: in proportion to total pop

i. Stratified Random: select random within substrata

j. Cluster: Selection within representative clusters

Collect the Data

k. Experiment: Control the environment

l. Observation:

2. Single Variable Data - Distributions

m. Graphing Categorical: Pie & bar chart)

n. Histogram (classes, count within each class)

o. – shape, center, spread. Symmetric, skewed right, skewed left

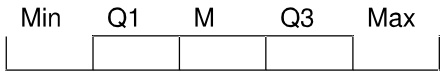
p. Stemplots

0	11222	0	112233
1	011333	0	56677
2	etc	1	

q. Mean: $\bar{x} = \sum x_i / n$

r. Median: M: If odd – center, if even - mean of 2

s. Boxplot:



t. Variance: $s^2 = \sum (x-\bar{x})^2 /(n-1) = SS_x /(n-1)$,

u. p78: standard deviation, s = $\sqrt{s^2}$

v. $SS_x = \sum (x-\bar{x})^2 = \sum x^2 - (\sum x)^2 / n$

w. Density curve – relative proportion within classes – area under curve = 1

x. Normal Distribution: 68, 95, 99.7 % within 1, 2, 3 std deviations.

y. p98: z-score $z = (x-\bar{x})/s$ or $(x-\mu)/\sigma$

z. Standard Normal: N(0,1) when N(μ,σ)

3. Bivariate - Scatterplots & Correlation

a. Explanatory – independent variable

b. Response – dependent variable

c. Scatterplot: form, direction, strength, outliers

d. – form is linear negative, …

e. – to add categorical use different color/symbol

f. p147: Linear Correlation- direction & strength of linear relationship

g. Pearsons Coeff: {-1 ≤ r ≤ 1} 1 is perfectly linear + slope, -1 is perfectly linear – slope.

h. $r = \frac{1}{n-1} * \sum \frac{(x-\bar{x})}{s_x}\frac{(y-\bar{y})}{s_y} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$,

i. $r = z_x z_y / (n-1)$,

j. $SS_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

4. Regression

k. least squares – sum of squares of vertical error minimized

l. p154: $y = b_0 + b_1 x$, or $\hat{y} = a + bx$,

m. (same as y = mx + b)

n. $b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{SS_{xy}}{SS_x}$ = r (s_y / s_x)

o. Then solving knowing lines thru centroid $((\bar{x},\bar{y}); a = \bar{y} - b\bar{x}$

p. $b_0 = \frac{\sum y - (b_1 \sum x)}{n}$

q. r^2 is proportion of variation described by linear relationship

r. residual = y - $\hat{y}$ = observed – predicted.

s. Outliers: in y direction -> large residuals, in x direction -> often influential to least squares line.
t. Extrapolation – predict beyond domain studied
u. Lurking variable
v. Association doesn't imply causation

5. Data – Sampling

a. Population: entire group
b. Sample: part of population we examine
c. Observation: measures but does not influence response
d. Experiment: treatments controlled & responses observed
e. Confounded variables (explanatory or lurking) when effects on response variable cannot be distinguished
f. Sampling types: Voluntary response – biased to opinionated, Convenience – easiest
g. Bias: systematically favors outcomes
h. Simple Random Sample (SRS): every set of n individuals has equal chance of being chosen
i. Probability sample: chosen by known probability
j. Stratified random: SRS within strata divisions
k. Response bias – lying/behavioral influence

6. Experiments

a. Subjects: individuals in experiment
b. Factors: explanatory variables in experiment
c. Treatment: combination of specific values for each factor
d. Placebo: treatment to nullify confounding factors
e. Double-blind: treatments unknown to subjects & individual investigators
f. Control Group: control effects of lurking variables
g. Completely Randomized design: subjects allocated randomly among treatments
h. Randomized comparative experiments: similar groups – nontreatment influences operate equally
i. Experimental design: control effects of lurking variables, randomize assignments, use enough subjects to reduce chance
j. Statistical signifi: observations rare by chance
k. Block design: randomization within a block of individuals with similarity (men vs women)

7. Probability & odds

a. 2 definitions:
b. 1) Experimental: Observed likelihood of a given outcome within an experiment
c. 2) Theoretical: Relative frequency/proportion of a given event given all possible outcomes (Sample Space)
d. Event: outcome of random phenomenon
e. n(S) – number of points in sample space
f. n(A) – number of points that belong to A
g. p 183: Empirical: P'(A) = n(A)/n = #observed/ #attempted.
h. p 185: Law of large numbers – Exp -> Theoret.
i. p. 194: Theoretical P(A) = n(A)/n(S) , favorable/possible
j. $0 \le P(A) \le 1$, $\sum$ (all outcomes) P(A) = 1
k. p. 189: S = Sample space, n(S) - # sample points. Represented as listing {(,), …}, tree diagram, or grid
l. p. 197 Complementary Events P(A) + P($\overline{A}$) = 1
m. p200: Mutually exclusive events**:** both can't happen at the same time
n. p203. Addition Rule: P(A or B) = P(A) + P(B) – *P(A and B) [which = 0 if exclusive]*
o. p207: *Independent Events:* Occurrence (or not) of A does not impact P(B) & visa versa.
p. Conditional Probability: P(A|B) – Probability of A given that B has occurred. P(B|A) – Probability of B given that A has occurred.
q. Independent Events iff P(A|B) = P(A) and P(B|A) = P(B)
r. Special Multiplication. Rule: P(A and B) = P(A)*P(B)
s. General mult. Rule: P(A and B) = P(A)*P(B|A) = P(B)*P(A|B)
t. <u>Odds / Permutations</u>
u. *Order important vs not* (Prob of picking four numbers)
v. *Permutations:* nPr, n!/(n – r)! , number of ways to pick r item(s) from n items if order is important : Note: with repetitions p alike and q alike = n!/p!q!.
w. *Combinations:* nCr, n!/((n – r)!r!) , number of ways to pick r item(s) from n items if order is NOT important
x. *Replacement vs not* (AAKKKQQJJJJ10) (a) Pick an A, replace, then pick a K. (b) Pick a K, keep it, pick another.
y. Fair odds - If odds are 1/1000 and 1000 payout. May take 3000 plays to win, may win after 200.

8. Probability Distribution

a. Refresh on Numb heads from tossing 3 coins. Do grid {HHH,….TTT} then #Heads vs frequency chart{(0,1), (1,3), (2,3), (4,1)} – Note Pascals triangle
b. Random variable – circle #Heads on graph above. "Assumes unique numerical value for each outcome in sample space of probability experiment".
c. Discrete – countable number
d. Continuous – Infinite possible values.

e. Probability Distribution: Add next to coins frequency chart a P(x) with 1/8, 3/8, 3/8, 1/8 values

f. Probability Function: Obey two properties of prob. ($0 \le P(A) \le 1$, $\sum$ (all outcomes) P(A) = 1.

g. Parameter: Unknown # describing population

h. Statistic: # computed from sample data

	Sample	Population
Mean	$\bar{x}$	μ - mu
Variance	s^2	σ^2
Standard deviation	s	σ - sigma

i. Base: $\bar{x} = \sum x/n$, $s^2 = \dfrac{\sum (x-\bar{x})^2}{(n-1)}$

	Frequency Dist	Probability Distribution
Mean	$\bar{x} = \sum xf / \sum f$	$\mu = \sum [xP(x)]$
Var	$s^2 = \dfrac{\sum (x-\bar{x})^2 f}{(\sum f - 1)}$	$\sigma^2 = \sum [(x-\mu)^2 P(x)]$
Std Dv	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

j. Probability acting as an $f / \sum f$. Lose the -1

9. Sampling Distribution

a. By law of large #'s, as n -> population, $\bar{x} \to \mu$

b. Given $\bar{x}$ as mean of SRS of size n, from pop with μ and σ. Mean of sampling distribution of $\bar{x}$ is μ and standard deviation is $\sigma / \sqrt{n}$

c. If individual observations have normal distribution N(μ,σ) – then $\bar{x}$ of n has N(μ, $\sigma / \sqrt{n}$)

d. Central Limit Theorem: Given SRS of b from a population with μ and σ. When n is large, the sample mean $\bar{x}$ is approx normal.

10. Binomial Distribution

a. Binomial Experiment. Emphasize Bi – two possible outcomes (success,failure). n repeated identical trials that have complementary P(success) + P(failure) = 1. binomial is count of successful trials where $0 \le x \le n$

b. p : probability of success of each observation

c. Binomial Coefficient: nCk = n!/(n – k)!k!

d. Binomial Prob: P(x = k) = $\binom{n}{k} p^k (1-p)^{n-k}$

e. Binomal μ = np

f. Binomal $\sigma = \sqrt{np(1-p)}$

11. Confidence Intervals

a. Statistical Inference: methods for inferring data about population from a sample

b. If $\bar{x}$ is unbiased, use to estimate μ

c. Confidence Interval: Estimate+/- error margin

d. Confidence Level C: probability interval captures true parameter value in repeated samples

e. Given SRS of n & normal population, C confidence interval for μ is: $\bar{x} \pm z^* \sigma / \sqrt{n}$

f. Sample size for desired margin of error – set +/- value above & solve for n.

12. Tests of significance

g. Assess evidence supporting a claim about popu.

h. Idea – outcome that would rarely happen if claim were true evidences claim is not true

i. Ho – Null hypothesis: test designed to assess evidence against Ho. Usually statement of no effect

j. Ha – alternative hypothesis about population parameter to null

k. Two sided: Ho: μ = 0, Ha: μ ≠ 0

l. P-value: probability, assuming Ho is true, that test statistic would be as or more extreme (smaller P-value is > evidence against Ho)

m. z = $\dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

n. Significance level α : if α = .05, then happens no more than 5% of time. "Results were significant (P < .01)"

o. Level α 2-sided test rejects Ho: μ = μ_o when u_o falls outside a level 1 – α confidence int.

a. Complicating factors: not complete SRS from population, multistage & many factor designs, outliers, non-normal distribution, σ unknown.

b. Under coverage and nonresponse often more serious than the random sampling error accounted for by confidence interval

c. Type I error: reject Ho when it's true – α gives probability of this error

d. Type II error: accept Ho when Ha is true

e. Power is 1 – probability of Type II error