AP Stat 01 Describe Data Cheat Sheet

by Jianmin Feng (taotao) via cheatography.com/79308/cs/20351/

Collect Data

type of variable	categorical vs quantitative (continuous vs discrete)
type of descriptive methods	tabular, graphic, Numberical
Tabular	n, f, rf, 100rf,cf, rcf, 100rcf
graphical	relations: bar, pie, dot, stem leave, histogram, cumulative freq
numberic	precise/inference, dull, complicated

graph the data

qualitative	bar pie
quanti-tative	dot plot,stemplot; histog-ram,cumulatice freq charts, boxplot
Examining graph	center(mean,median,mode), spread (range, std, variance), shape (symmetric, skewed)
pattern/d-eviations	cluster/gap, outliers
dotplot	spread,shape,approx center
stemplot	shaped,spread, center
histograph	f vs rf, shape/center, large dataset, error bar for spread
cumulative freq charts	S shaped, T(skewed) shape, meaningful order

Central tendency - mean, median, mode

summering distribution	population/sample, center/spread/sape
mean	mu=population mean,X bar=sample mean

Central tendency - mean, median, mode (cont)

median	for skewed data, odd/even sample size
mode	number with highest freq
symmet-rical	mean=median=mode
left skewed	mode>median>mean
right skewed	mode>mean>media

variance/spread - range, IQR, STD

variance	spread from mean
range	largest-smallest measur-ement, outliers affect
IQR	interquartile range, eg Q3-Q1, not affected by outliers, median / IQR
STD	standard deviation, square root of variance, outlier affect, >=0
variance	average the square of deviation from mean
population variance	N, sigma, mu
sample variance	n-1, x bar, s

mean/STD, median/IQR

Position - quartile,percentile,standarded score

percentile	order, divide into 100 equal parts, count kth perncentile Pk
quartile	order, divide into 4 equal parts (median calc), count kth quartile Qk, P25=Q1, P50=Q2..
z score	standardized score, (x-mea-n)/std, compare datasets with different scales, eg temper-ature in north vs south city

Graphing uni variant data

graphical summaries	Y scale:misleading manipu-lation
box plots	box(Q2-Q3) and whiskers(-lower Q1,upper Q4), whiskers <1.5IQR (Q3-Q1), L=Q1-1.5IQR, U=Q3+1.5IQR. point >U or <L are outliers
	based on position, identify outlier and general shape(-skewed or not)
	calc: Q1, Median, Q3, IQR, L, U
shift unit +a	(variance/spread) range, std,IQR not affected
enlarge or shrink unit,*b	all stat enlarged or shrinked
Compare distri-butions	center, spread, shape
	outerlier or unusual values,cl-uster/gap
	context of the question
	dot plot, stemplot, histogram, freq polygram

Avoid simple list the stat (center,std and shape), instead, make a clear comparative statement.

Bivariant data

Scatter plot	shape: linear, non-linear, no relation
	direction: positive or negative linear relation
	strength of linear relation: close to the line

Numeric methods

correl-ation coeffi-cient	degree and direction of linear relation of two quantitative variables (x,y)
	rho and r, [-1,+1]

By **Jianmin Feng** (taotao)
cheatography.com/taotao/

Not published yet.
Last updated 17th December, 2019.

Bivariant data (cont)

	0, 0.1, 0.5,0.85,1
least squares regression line	
formular	Y = a+ bX +e
Y	dependent/response variable
x	independent/explanatory variable
a	y intercept of line
b	slope of the line
e	random error, residual error
predicted vaue	y hat
residual error	e
least square regression	minimize the sum of squares of residual error
	line of best fit (X bar, Y bar), slope=r(Sy/Sx)
coefficient of determ-ination	R squared, percent of variance of Y determined by variance X
	[-1,+1]
influential point	point that affect the correl-ation efficient
Outlier	maybe influential point
residual plot	should be random, or else, fit is not the best
transform-ation to fit linear	log, sqrt,reciprocal,squar-e,power

1 calc slope, intercept, write formula, plot the linear line
2 make a prediction, calc residual error
3 calc coefficient of determination r = SSxy/sqrt(SSxx * SSyy)

stat and interpretation

categorical data

marginal and joint freq of two way tables	
contingency(-joint) table	r*c
marginal	row col grant total
conditional relative frequency	
association	compare with row total * col total /grand total