INTRODUCTION: STATISTICS IS THE SCIENCE OF (1) OBTAINING DATA,

(2) SUMMARIZING DATA, AND

(3) USING DATA TO MAKE IMPORTANT INFERENCES. (STATISTICAL/PROBALISTIC)

(1) OBTAINING DATA.

(a) DESIGNED EXPERIMENT: "A RANDOMIZED CONTROLLED" EXPERIMENT THAT IS SPECIFICALLY DESIGNED TO EXAMINE A SPECIFIC THING(S)

Exs: SALK VACCINE; BLOCK STUDIES (METHODS, COLA)

(b) OBSERVATIONAL STUDY: OBSERVATION OF RELATIONSHIPS WITH SUBSEQUENT ANALYSIS TO ADJUST FOR CONFOUNDING FACTORS.

Exs: EPA; left-handedness; human smoking

(c) SAMPLING: UTILIZING GOOD RANDOM SELECTION FROM POPULATION

Exs. Literary Digest; ESPN call in-poll; Gallup poll.

(d) PUBLISHED REPORTS: WHICH ARE HOPEFULLY EITHER GOOD STUDIES USING (a), (b), OR (c), ABOVE, OR ARE A COMPLETE (CENSUS) OF USEFUL INFORMATION.

(2) SUMMARIZATION OF DATA.

- BASIC DESCRIPTIVE STATISTICS (D.S.) ON EACH VARIABLE
  THESE D.S. INCLUDE MEASURES OF CENTER,
  LIKE AVERAGE, AND MEDIAN;
  MEASURES OF SPREAD, LIKE STANDARD DEVIATION
  AND INTERQUARTILE RANGE

- PICTURES/GRAPHIC ON EACH VARIABLE,
  LIKE HISTOGRAM, STEM/LEAF CHART, BOX PLOT, ETC

- SUMMARY STATISTICS LOOKING AT RELATIONSHIP
  BETWEEN VARIABLES, LIKE RELATIVE %'s.

- SUMMARY PICTURES / GRAPHICS
  COMPARING VARIABLES, LIKE X, Y (VAR. VAR) PLOTS,
  ETC.

(3) USE SCIENCE OF PROBABILITY AND
RANDOM VARIABLES TO MAKE
INFERENCES LIKE CONFIDENCE
INVERVALS AND TESTS OF HYPOTHESES
ABOUNT POPULATION PARAMETERS.

DESCRIPTIVE STATISTICS

QUALITATIVE RESPONSE: (YES/NO); (ELECTRIC, GAS, OIL)

e.g. DID Child get Polio — Yes / No

HOUSE'S MAIN HEATING SOURCE — ELECTRIC / GAS / OIL

USE BAR CHARTS OR PIE CHARTS TO SHOW PROPORTIONS (p's) OR p·100% = %'s.

QUANTITATIVE RESPONSE: HEIGHT, WEIGHT, SCORE, ETC.

- USE HISTOGRAM OR STEM/LEAF CHART TO SHOW DISTRIBUTION OF VALUES FOR A VARIABLE
- USE BOXPLOT TO VISUALLY SUMMARIZE KEY SUMMARIES & LOOK FOR OUTLYING OBSERVATIONS FOR A VARIABLE
- USE PLOTS TO LOOK AT RELATIONSHIP BETWEEN VARIABLES

(P): PROBABILITY: CHANCE ON A SCALE FROM [0.0 to 0.5 to 1.0]

can't happen | Equally likely to happen vs not happen | will happen

Ex 1: A fair coin is tossed once.

P(get a head) = P(don't get head) so P(get head) = 0.5 = $\frac{1}{2}$

[In very many repeats of Ex 1, about $\frac{1}{2}$ of time you'd get head; $1-\frac{1}{2}=\frac{1}{2}$ time you'd NOT get a head

Ex 2. ROLL A FAIR DIE ONCE.

$$\boxed{\underbrace{P(\text{get a } \boxed{2})}_{P(\boxed{2})}} = \frac{1}{6} = \boxed{1 - P(\text{don't get a } \boxed{2})} = 1 - \frac{1}{6} = \frac{5}{6}.$$

GENERAL RULE 1: FOR ANY EVENT A, P(A) = 1 − P(not A)

same as: P(not A) = 1 − P(A)

In Ex 2, P(get an even number) = $\frac{3}{6}$,

because there are a total of 6 possible outcomes — specific things that can occur — of the roll, namely, {1, 2, 3, 4, 5, 6}

Each of these are equally likely, and

For 3 of the 6, an even number occurs.

GENERAL RULE 2: IF ALL THE POSSIBLE OUTCOMES ARE EQUALLY LIKELY, THEN FOR ANY EVENT A,

$$P(A) = \frac{\#\text{ possible outcomes so that A occurs}}{\text{total }\#\text{ possible outcomes}}$$

Box: $\boxed{R1}\ \boxed{R2}\ \boxed{B1}\ \boxed{B2}$

Ex 3. A box has 2 red cards: R1, R2 and 2 blue cards: B1, B2.

A hand of 2 cards are drawn (without replacement) from the box.

NOTE there are 4·3 = 12 possible outcomes of expt: namely,

(R1,R2), (R2,R1), (R1,B1), (R1,B2), (B1,R1), (B2,R1), (R2,B1), (B1,R2), (R2,B2), (B2,R2), (B1,B2), (B2,B1)

so, P(get red on 2^{ND} draw) = $\frac{6}{12}$; P(red on 1^{ST} and red on 2^{ND}) = $\frac{2}{12}$

NOW, IF YOU KNEW A RED WAS DRAWN ON 1^{ST} DRAW, THEN THERE IS ONLY ~~A~~ 1 in 3 (OR $\frac{1}{3}$ OR $\frac{2}{6}$) chance that a RED IS DRAWN ON 2^{ND},

so, {P(RED ON 2^{ND} | RED ON 1^{ST}) = $\frac{1}{3} = \frac{2}{6}$}, this in a conditional probability.

("|" = given that got)

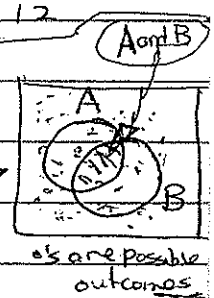
A CONDITIONAL PROBABILITY IS ALSO A PROBABILITY, ITS JUST UNDER THE PARTICULAR CONDITION.

In last ex., e.g. $P\left(\text{red on } 2^{ND} \,\middle|\, \text{red on } 1^{ST}\right) = 1 - P\left(\text{NOT red on } 2^{ND} \,\middle|\, \text{red on } 1^{ST}\right) = 1 - \frac{2}{3} = \frac{1}{3}.$

CALVIN & HOBBS

Continuing last ex.,

we see $\frac{2}{12} = P\left(\text{red on } 1^{ST} \text{ and red on } 2^{ND}\right) = \underbrace{P\left(\text{red on } 1^{ST}\right)}_{\frac{6}{12}} \cdot \underbrace{P\left(\text{red on } 2^{ND} \,\middle|\, \text{red on } 1^{ST}\right)}_{\frac{1}{3}} = \frac{2}{12}$



GENERAL RULE 3: FOR ANY 2 EVENTS A and B,

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

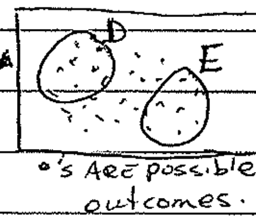
BOX: $\boxed{R1}\boxed{R2}\boxed{G1}\boxed{B1}$

NEW $\boxed{\text{Ex. 5}}$ CONSIDER A BOX WITH 2 REDS: R1, R2 ; 1 GREEN: G1 and 1 Blue: B1

ONE DRAW IS MADE FROM BOX,

$P\left(\text{get GREEN on DRAW OR get BLUE on DRAW}\right) = \frac{2}{4} = P\left(\text{get G1 or get B1}\right) = P(\text{get G1}) + P(\text{get B1}) = \frac{1}{4} + \frac{1}{4} = \frac{2}{4}$

GENERAL RULE 4: IF 2 events D and E are mutually exclusive (one event occurring excludes other event from occurring), then $P(D \text{ or } E) = P(D) + P(E)$.



NOTE: GENERAL RULE 4 EASILY EXTENDS TO MORE THAN 2 EVENTS, e.g. IF D1, D2, D3, ARE MUTUALLY EXCLUSIVE THEN $P(D1 \text{ or } D2 \text{ or } D3) = P(D1) + P(D2) + P(D3)$

ALSO, GENERAL RULE 3 EASILY EXTENDS TO MORE THAN 2 EVENTS, e.g. FOR ANY 3 EVENTS A1, A2, A3; $P(A1 \text{ and } A2 \text{ and } A3) = P(A1) \cdot P(A2|A1) \cdot P(A3|(A1 \text{ and } A2))$

$\boxed{\text{Ex. 6}}$ Standard Deck of 52 cards (13 Hearts, 13 Spades, 13 Clubs, 13 Diamonds, & for each of these suits have 2, 3, ..., 10, J, Q, K, A)

A hand of 3 cards is dealt (without repl.)

What's P(get exactly 2 Q's in hand)

1st, 2nd, 3rd:

Q and Q and Not Q: $\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{48}{50}$ ← Rule 3 extended

OR +

Q — Not Q — Q: $= \frac{4}{52} \cdot \frac{48}{51} \cdot \frac{3}{50}$ ← Rule 4 extended

OR +

Not Q — Q — Q: $\frac{48}{52} \cdot \frac{4}{51} \cdot \frac{3}{50}$

Ex. 6 b

NOTE: If instead had dealt 3 cards $\boxed{\text{WITH}}$ replacement from deck then P(exactly 2 Q's in 3 independent cards) =

$$\left.\begin{matrix} \frac{4}{52} \cdot \frac{4}{52} \cdot \frac{48}{52} \\ + \\ \frac{4}{52} \cdot \frac{48}{52} \cdot \frac{4}{52} \\ + \\ \frac{48}{52} \cdot \frac{4}{52} \cdot \frac{4}{52} \end{matrix}\right\} = \boxed{3\left[\left(\frac{4}{52}\right)^2\left(\frac{48}{52}\right)\right]}$$

$$\boxed{= \binom{3}{2}\left[\left(\frac{4}{52}\right)^2\left(1-\frac{4}{52}\right)^{(3-2)}\right]}$$

In general two trials are independent, if no matter what you get on the 1st trial, that it has no influence on probs for 2nd trial, so under WITH replacement, the 3 dealt cards are independent of one another. (under WITHOUT replacement, they are NOT indpt.)

* NOTE: If 2 trials are indpt & E_1 from 1st trial & E_2 from 2nd trial, then $P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2 | E_1) \overset{\text{indpt}}{=} P(E_1)P(E_2)$

GENERAL RULE 5 for independent trials: "Binomial Rule"

$$P\left(\begin{matrix}\text{particular "thing" occurs} \\ \text{on exactly } k \text{ out of } n \\ \text{(independent) trials}\end{matrix}\right) = \binom{n}{k} \cdot \left[p^k (1-p)^{(n-k)}\right]$$

where p is chance the thing occurs on any given trial and

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ = # of ways to choose the k trials out of the n total trials for which the thing occurs.

NOTE: $n! = n(n-1)(n-2)\cdots(2)(1)$

NOTE: our ex. 6 b above, is an illustration of General Rule 5, with $n = 3$; "thing" is Q; $k = 2$ and $p = 4/52$

NOTE 2: If roll a fair die 10 (indpt) times, P(get $\boxed{2}$ on exactly 4 of 10 rolls) $= \binom{10}{4}\left[\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6\right]$

Discrete Random Variables

Ex.7 A fair coin is flipped two (indpt) times.

Let the random variable X be the $\underbrace{\text{number of heads in the 2 flips}}_{\#}$

Let's find the "probability distribution" for X.

Prob.	Possible Outcomes of 2 Tosses (1st and 2nd)	X
$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	(T, T)	0
$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	(T, H)	1
$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	(H, T)	1
$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	(H, H)	2

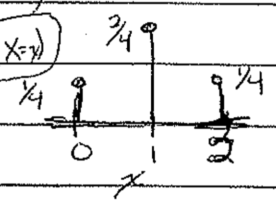
GR. 3

so, $P(X=0)=1/4$

GR. 4 $\rightarrow$ $P(X=1)=1/4+1/4=2/4$

$P(X=2)=1/4$

$\Rightarrow$ $P(X=x)$ Also labelled $f(x)$



Here, NOTE: $P(X \geq 1) = P(X=1) + P(X=2) = 2/4 + 1/4 = 3/4$.

GENERAL RULE 7: FOR ANY DISCRETE R.V. X AND ANY INTERVAL "I" OF REAL LINE

$$P(X \text{ is in } I) = \sum_{\substack{\text{all possible} \\ x \text{ in } I}} P(X=x) \text{ also NOTED as } \sum_{\substack{\text{all possible} \\ x \text{ in } I}} f(x)$$

GOING BACK TO EX. 7, NOTE that the long-run average value that X would take is:

$0\left(\frac{1}{4}\right) + 1\left(\frac{2}{4}\right) + 2\left(\frac{1}{4}\right)$, since probs are long-run fractions

here equals $0 + 2/4 + 2/4 = 1$

The long-run average value for X is also labelled as the mean of X, and also called the expectation for X and also called expected value for X and written as $E(X)$; also written as μ

General Rule 8: For any discrete r.v. X, $\boxed{E(X) = \sum_{\text{all possible } x} x f(x) \equiv \mu}$

Note: $E(X)$ is a measure of the center of the distribution of X.

The $\boxed{\text{Variance}(X)} \equiv Var(X) = E((X-\mu)^2)$ = long-run average of the squared distance from μ. $\equiv \sigma^2$

It is a measure of the spread of the distribution of X.

$\boxed{\text{Standard Deviation}(X)} = SD(X) = \sqrt{Var(X)} \equiv \sigma$

GOING BACK TO EX 7 (# of heads in 2 TOSSES : X).

Recall $\mu = 1$, so

$$\sigma^2 = Var(X) = E((X-\mu)^2) = E((X-1)^2) = \sum_{\text{all poss. } x} (x-1)^2 P(X=x) = \sum_{\text{all poss } x} (x-1)^2 f(x)$$

$$= (0-1)^2(\tfrac{1}{4}) + (1-1)^2(\tfrac{2}{4}) + (2-1)^2(\tfrac{1}{4}) = \tfrac{1}{4} + 0 + \tfrac{1}{4} = \tfrac{1}{2}$$

so, $\sigma = SD(x) = \sqrt{Var(x)} = \sqrt{\tfrac{1}{2}} \doteq .71$

roughly speaking the SD(x) is "typical" distance X is away from μ in long-run.

Ex. 8 Say Tom spends \$1 on a lottery ticket, he has a $\frac{1}{1000}$ chance of winning $\boxed{\$100 \text{ jackpot}}$, otherwise he wins nothing (no jackpot win). What's E(Tom's net gain), What's StdDev(Tom's net gain)? Tom's net gain is a r.v., call it X.

Note: $P(X = (-1 + 100)) = \frac{1}{1000}$ while $P(X=-1) = \frac{999}{1000}$

99, -1: cost of ticket

so $E(X) = 99(\frac{1}{1000}) + (-1)\frac{999}{1000} = -0.9$

while $SD(x) = \sqrt{(99-(-0.9))^2(\frac{1}{1000}) + (-1-(-0.9))^2(\frac{999}{1000})}$

One important "type" of Discrete r.v. is Binomial (n, p):

It is as follows:

Let X be the # of times a particular "thing" occurs in n independent trials, where there is p chance it will occur on given trial,

by GENERAL RULE 5:

for $K = 0, 1, 2, \ldots, n$, $P(X = K) = \binom{n}{K} \cdot [p^K (1-p)^{(n-K)}]$

Ex. 9

An illustration of this is: → so indpt. draws.

Make 3 draws WITH replacement from box that has 6 reds and 4 non-red tickets in it, i.e. $\boxed{6 \text{ R's} \quad 4 \text{ N's}}$

Let r.v. X be $\boxed{\text{\# of reds drawn}}$, then $X \sim \text{binomial}\left(n=3,\ p=\frac{6}{6+4}\right)$

with for $x=0,1,2,3$, $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{3}{x}\left(\frac{6}{10}\right)^x \left(\frac{4}{10}\right)^{(3-x)}$

NEW DISTRIBUTION.

Ex. 10 Now go back to Ex. 9, but have the 3 draws instead be drawn WITHOUT replacement from the same box $\boxed{6 \text{ R's} \ 4 \text{ N's}}$

Let r.v. X be # of reds drawn.

then for $x=0,1,2,3$, $P(X=x) = \dfrac{\binom{6}{x}\binom{4}{3-x}}{\binom{10}{3}} \equiv f(x)$

$X \sim$ hypergeometric$(3,6,4)$

The # OF OCCURRENCES IN AN "SELF-SIMILAR" INTERVAL OF TIME OFTEN $\approx$ HAS A POISSON(λ) DISTRIBUTION, WHERE λ IS the MEAN VALUE.

Ex. 11 # OF TRAFFIC ACCIDENTS ON SILAS CREEK PARKWAY FOR A MONTH'S RUSH HOURS. $\tilde{\sim}$ POISSON(λ)

for $x=0,1,2,3,\ldots$, $P(X=x) = \dfrac{e^{-\lambda}\lambda^x}{x!} = f(x)$

Ex. 12 : # OF PEOPLE SERVED @ LOCAL McD's FROM 2-5 PM $\tilde{\sim}$ Poiss.

CONTINUOUS RANDOM VARIABLES.

Continuous R.V's take values on a "continuum" (e.g. on an interval)

$\boxed{\text{Ex. 12}}$ For ex. (a) height of randomly selected US adult male might take any value in the interval (48.0", 86.5");

(b) randomly chosen point (position) on a yardstick could be any value in the interval (0.0", 36.0").

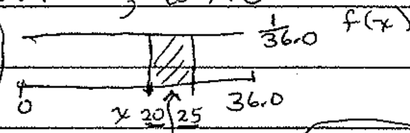
(c) yield on a cotton plot might be any value in (20.0, 90.0 $\frac{bu}{ac}$).

(d) amount of time until the next traffic accident occurs.

$\boxed{\text{UNIFORM}}$ One Major "FAMILY" OF Continuous R.V's is UNIFORM (a, b)

The ex 12(b), above, randomly chosen point (position) on yardstick $\equiv X$.

Might have a Uniform(0.0", 36.0") distribution, with relative likelihood, density "f", looking like a flat line



To get $P(X \text{ is in } (20.0, 25.0))$ = Area under f(x) between 20 and 25 (shaded above)

which here = $\frac{5}{36}$.

NOTE: $P(X \text{ is in } (20, 25)) = \boxed{\int_{20}^{25} f(x)\,dx}$

NOTE: for continuous r.v, f is a "density" not a probability

$\boxed{\text{GENERAL RULE 9}}$: FOR ANY CONTINUOUS R.V. X, WITH DENSITY, f,

$P(X \text{ is in interval } A) = \int_{x \text{ in } A} f(x)\,dx$ = Area under f over region A.

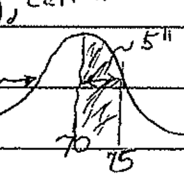
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$\sigma^2 = \boxed{Var(X) = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx}$ $\quad$ $StdDev(X) = \sqrt{Var(X)} = \sigma$

ANOTHER MAJOR "FAMILY" IS NORMAL(μ, σ)

$\boxed{\text{NORMAL}}$ FOR ex 12(a) for male ht. might be $\sim$ Normal(μ = 70", σ = 5"), call it X

with $P(X \text{ is in } (70, 75))$ is the shaded Area $= \int_{70}^{75} f(x)dx$ with f(x)



To approx. this area, 1ST convert (standardize) the endpoints X's

$\boxed{Z = \frac{X-\mu}{\sigma}}$ AND then use the standard normal table:

here $P(X \text{ is in } (70, 75)) = P(Z \text{ is in } (\frac{70-\mu}{\sigma}, \frac{75-\mu}{\sigma})) = P(Z \text{ is in } (0, 1)) \doteq .84 - .50 = .34$

$P(Z \le 1.0) - P(Z \le 0.0)$

using (std) normal table

CONTINUOUS RANDOM VARIABLES (CONT.)

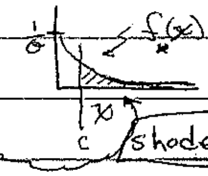
ANOTHER MAJOR FAMILY IS EXPONENTIAL(θ)

EXPONENTIAL For 12d, time until next accident occurs, call it X often has Exp(θ) distribution, where θ is mean of dist.

the density $f(x) = \frac{1}{\theta} e^{-x/\theta}$ for $x > 0$.

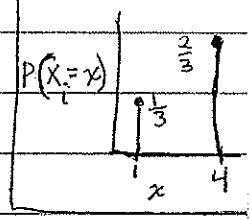
For any constant $c > 0$

$P(\text{time exceeded "c"}) = P(X > c) = \int_c^\infty \underbrace{\frac{1}{\theta} e^{-x/\theta}}_{f(x)} dx = e^{-c/\theta} =$ shaded area



SAMPLING DISTRIBUTION

Ex. 13 CONSIDER A VERY SIMPLE DISCRETE DISTRIBUTION ("r.v. X"), where possible values for X are either "1" OR "4" AND $P(X=1)=\frac{1}{3}$; $P(X=4)=\frac{2}{3}$. $\Rightarrow$



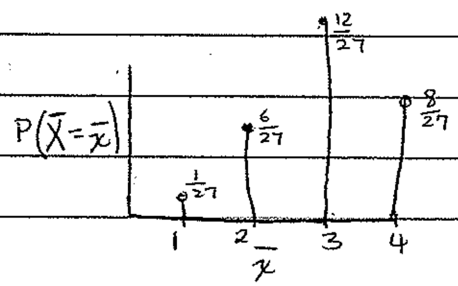
You CAN VIEW THIS AS A POPULATION OF VALUES, WHERE $\frac{1}{3}$ OF THE VALUES ARE "1" AND $\frac{2}{3}$ OF THE VALUES ARE "4"

AND THINK OF X AS A RANDOM DRAW FROM POPULATION

NOW SUPPOSE THAT $n=3$ INDEPENDENT (WITH REPLACEM.) RANDOM DRAWS (X_1, X_2, X_3) ARE MADE FROM POPULATION. Then Note that the possible VALUES FOR the (X_1, X_2, X_3) triples (samples), their probabilities, and "sample statistic $\bar{X}$" ARE:

prob.	possible (X_1, X_2, X_3)	Sample average $\bar{X}=\frac{1}{n}\sum X_i$	
$\frac{1}{27}=(\frac{1}{3})(\frac{1}{3})(\frac{1}{3})$	(1, 1, 1)	$\frac{1}{3}(1+1+1)=1$	NOTE! $P(\bar{X}=1)=\frac{1}{27}$
$\frac{2}{27}=\frac{2}{3}\,\frac{1}{3}\,\frac{1}{3}$	(4, 1, 1)	$\frac{1}{3}(4+1+1)=2$	NOTE: $P(\bar{X}=2)=\frac{2}{27}+\frac{2}{27}+\frac{2}{27}=\frac{6}{27}$
$\frac{2}{27}=\frac{1}{3}\,\frac{2}{3}\,\frac{1}{3}$	(1, 4, 1)	$\frac{1}{3}(1+4+1)=2$	
$\frac{2}{27}=\frac{1}{3}\,\frac{1}{3}\,\frac{2}{3}$	(1, 1, 4)	$\frac{1}{3}(1+1+4)=2$	
$\frac{4}{27}=\frac{2}{3}\,\frac{2}{3}\,\frac{1}{3}$	(4, 4, 1)	$\frac{1}{3}(4+4+1)=3$	NOTE: $P(\bar{X}=3)=\frac{4}{27}+\frac{4}{27}+\frac{4}{27}=\frac{12}{27}$
$\frac{4}{27}=\frac{2}{3}\,\frac{1}{3}\,\frac{2}{3}$	(4, 1, 4)	$\frac{1}{3}(4+1+4)=3$	
$\frac{4}{27}=\frac{1}{3}\,\frac{2}{3}\,\frac{2}{3}$	(1, 4, 4)	$\frac{1}{3}(1+4+4)=3$	
$\frac{8}{27}=\frac{2}{3}\,\frac{2}{3}\,\frac{2}{3}$	(4, 4, 4)	$\frac{1}{3}(4+4+4)=4$	NOTE: $P(\bar{X}=4)=\frac{8}{27}$

$\Rightarrow$



Ex.13 (continued).

NOTE: $\underbrace{E(\bar{X})}_{\text{label as } \mu_{\bar{X}}} = \sum_{\substack{\text{all possible}\\ \bar{x}}} \bar{x}\, P(\bar{X}=\bar{x}) = 1\left(\frac{1}{27}\right) + 2\left(\frac{6}{27}\right) + 3\left(\frac{12}{27}\right) + 4\left(\frac{8}{27}\right) = 3.00$

Also NOTE that the mean μ of population:

$\mu_{X_i} = E(X_i) = 1\left(\frac{1}{3}\right) + 4\left(\frac{2}{3}\right)$, which also $= 3.00$

SINCE $E(\bar{X}) = \mu$, $\bar{X}$ is AN UNBIASED ESTIMATOR OF μ.

NOTE: $\text{StdDev}(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \sqrt{E\left[(\bar{X}-\mu_{\bar{X}})^2\right]} \equiv \sigma_{\bar{X}}$

$= \sqrt{\sum_{\bar{x}} (\bar{x}-3)^2 P(\bar{X}=\bar{x})} = \frac{\sqrt{2}}{\sqrt{3}}$

ALSO NOTE that the StdDev σ OF population

$\sigma_{X_i} = \sqrt{\text{Var}(X_i)} = \sqrt{E\left[(X_i-\mu_{X_i})^2\right]} = \sqrt{2}$

NOTE: here, $\sigma_{\bar{X}} = \frac{\sigma_{X_i}}{\sqrt{n}}$ AND $\mu_{\bar{X}} = \mu_{X_i}$.

$\boxed{\text{GENERAL RULE 10}}$ FOR n indpt draws from popul with mean μ, SD σ,

$\bar{X} \equiv \left[\frac{1}{n} \sum X_i\right]$ ("sample average")

$E(\bar{X}) = \mu$; $\underbrace{\text{StdDev}}_{\text{SD}}(\bar{X}) = \frac{\text{Std Dev}(X_i)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$

CENTRAL LIMIT THEOREM (CLT).

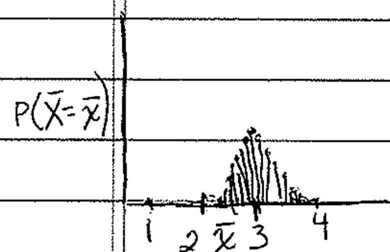
GOING BACK TO OUR LAST EXs (EX 13).

BUT NOW INSTEAD WE SAMPLED $n=100$

DRAWS FROM THE (SAME) POPULATION

THEN (BY G.R. 10), $E(\bar{X}_{100}) = \mu = 3$ AND $SD(\bar{X}_{100}) = \frac{\sigma}{\sqrt{100}} = \frac{\sqrt{2}}{10} = .1414$

NOW, since n large, the distrib. of $\bar{X}$ will be close to being symmetric and bell-shaped, like a normal curve and, for this ex., the distrib. of $\bar{X} \approx$



i.e. $\bar{X}$ is Approx. Distrib. Normal$\left(\mu_{\bar{X}} = \mu_X = 3;\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = .1414\right)$

THIS IS AN ILLUSTRATION OF THE CENTRAL LIMIT THEOREM.

$\boxed{\text{GENERAL RULE 11}}$ FOR LARGE n SAMPLE FROM POPULATION (μ, σ)

$\boxed{\text{CENTRAL LIMIT THM (CLT)}}$

$\bar{X} \approx \text{normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, NO MATTER THE POPUL. DIST.

NOTE: WE CAN USE CLT, TO GET APPROX. PROBs RELATING TO $\bar{X}$ WHEN n LARGE

in last ex, $n=100$; $\mu=3$; $\sigma=\sqrt{2} \doteq 1.414$

$$P(\bar{X}_{100} \le 3.1414) = P\left(\frac{\bar{X}_{100}-3}{\frac{\sqrt{2}}{\sqrt{100}}} \le \frac{3.1414-3}{.1414}\right) \doteq P(Z \le 1) \doteq .84$$

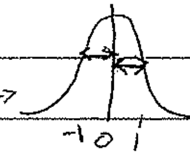
std normal

from normal table

LARGE SAMPLE CONFIDENCE INTERVALS FOR μ

AGAIN, n large, $X_1, X_2, \ldots, X_n$ (indpt) drawn from population with mean μ & SD σ.

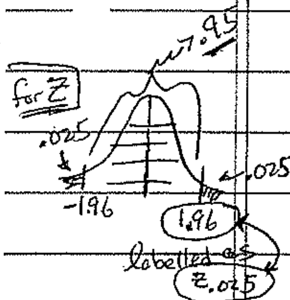
$$\bar{X}_n \approx \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \Rightarrow \boxed{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}} \;\dot{\sim}\; \underbrace{Z}_{\text{std normal}} \rightsquigarrow$$



ALSO, if n large, then sample SD, $s = \sqrt{\frac{1}{n}\left(\sum (x_i - \bar{X})^2\right)}$ is "very close to σ",

So $\boxed{\frac{\bar{X}_n - \mu}{s/\sqrt{n}}} \;\dot{\sim}\; Z$. From Z table, $\boxed{P(Z > 1.96) = 1 - P(Z \le 1.96) = 1 - .975 = .025}$

ALSO $P(Z < -1.96) = .025$



So, $\boxed{P\left(-1.96 \le \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \le 1.96\right)} \doteq P(-1.96 \le Z \le 1.96) = 1 - 2(.025) = .95$

$$\Updownarrow$$

$\boxed{P\left(-1.96 \frac{s}{\sqrt{n}} \le \bar{X}_n - \mu \le 1.96 \frac{s}{\sqrt{n}}\right)} = .95$

$$\Updownarrow$$

$\boxed{P\left(\bar{X} - 1.96 \frac{s}{\sqrt{n}} \le \mu \le \bar{X} + 1.96 \frac{s}{\sqrt{n}}\right)} = .95$

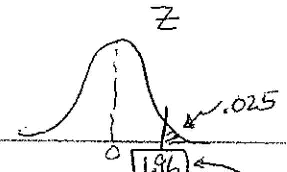
$$\Updownarrow$$

$$P\left(\underline{\underline{\mu}} \text{ is in interval } \left(\bar{X} - 1.96\, s/\sqrt{n},\ \bar{X} + 1.96\, s/\sqrt{n}\right)\right) = .95$$

from the sample (of size n), you get an observed sample avg $\bar{x}$ $\boxed{= \frac{1}{n}\left(\sum x_i\right)}$ and an observed $s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$,

and a $(.95) \times 100\% = 95\%$ confidence interval (C.I.) for μ is:

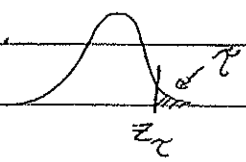
$$\boxed{\left(\bar{x} - 1.96 \frac{s}{\sqrt{n}},\ \bar{x} + 1.96\, s/\sqrt{n}\right)}$$

RECALL: 

$\boxed{1.96} = \boxed{Z_{.025}}$ since .025 of Area under Z curve is to right of 1.96

(NOTE: $\boxed{1.96}$ is ALSO the $1-.025$ quantile or 97.5^{th} percentile of Z.)

IN GENERAL, $\boxed{Z_\tau}$ is value so that area under Z curve to the right of it is τ.



GENERAL RULE 12. FOR LARGE n, A $\boxed{(1-\alpha)*100\%\ \text{C.I for } \mu}$ is:

$$\boxed{\left(\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + Z_{\alpha/2}\,{}^{S}\!/\!{\sqrt{n}}\right)}$$

Also written as:

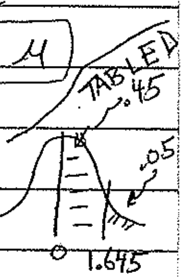
$$\boxed{\bar{x} \pm Z_{\alpha/2}\,{}^{S}\!/\!{\sqrt{n}}}$$

Check: A $\boxed{95\%\ \text{CI for } \mu} = \boxed{(1-\alpha)*100\%\ \text{CI for } \mu} \Rightarrow \alpha = .05$

$\Rightarrow \alpha/2 = .025$; To GET $Z_{\alpha/2} = Z_{.025} \Rightarrow 1.96$

$\Rightarrow$ 95% CI for μ of: $\boxed{\bar{x} \pm 1.96\,{}^{S}\!/\!{\sqrt{n}}}$, as before

Ex. 14 A sample of $n=400$ is taken from very large popul. the observed sample average is 23.2, while the Observed SD is 8.0. Then, construct a $\boxed{90\%\ \text{CI. for } \mu}$



since $90\% = (1-\alpha)*100\% \Rightarrow \alpha = .10 \Rightarrow \alpha/2 = .05$ AND

so $Z_{\alpha/2} = Z_{.05} = 1.645$

SHOW DEMONSTRATIONS

so 95% CI for μ is: $\boxed{23.2 \pm 1.645\left(\frac{8.0}{\sqrt{400}}\right)} \Rightarrow \boxed{23.2 \pm 0.65}$

GENERAL RULE

$\underline{X} \equiv (X_1, X_2, \ldots, X_n)$.

GENERAL RULE FOR (ANY) PARAMETER, SAY τ.

Let $u_1(\underline{X})$ and $u_2(\underline{X})$ be two functions of sample and constants so that:

$$P(u_1(\underline{X}) \le \tau \le u_2(\underline{X})) = 1-\alpha,$$

THEN $(u_1(\underline{x}), u_2(\underline{x}))$ is a $(1-\alpha) * 100\%$ CI for τ.

Ex. SAY X_1 is a draw from a Normal$(\mu, \sigma = 3)$.

CONSTRUCT A 95% CI for μ, based only on X_1 & constants.

NOTE: $\frac{X_1 - \mu}{3} \sim Z$ (std normal).

so, $P\left(-1.96 \le \frac{X_1 - \mu}{3} \le \underbrace{1.96}_{\cancel{z_{\frac{1-.95}{2}}}}\right) = .95$

$$\Rightarrow P(\underbrace{X_1 - 1.96(3)}_{u_2(\underline{X})} \le \underbrace{\mu}_{\tau} \le \underbrace{X_1 + 1.96(3)}_{u_1(\underline{X})}) = .95$$

so, $(x_1 - 1.96(3), x_1 + 1.96(3))$ is a 95% CI for μ

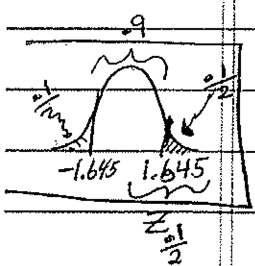
TEST OF HYPOTHESIS:

SAMPLE FROM SINGLE POPULATION.

n LARGE: TEST ON μ

FOR Ex. 14, WE HAD A SAMPLE OF $N = 400$ WITH SAMPLE AVG, $\bar{x} = 23.2$ AND SAMPLE SD, $S = 8.0$.

& WE FORMED A 90% [i.e. $(1 - \overset{\alpha}{.1}) * 100\%$] C.I FOR μ



BY $\boxed{\bar{x} \pm Z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}} \Rightarrow \boxed{\bar{x} \pm 1.645 \left(\frac{8.0}{\sqrt{400}}\right)} \Rightarrow$

$\Rightarrow \boxed{\bar{x} \pm 0.65} \Rightarrow \boxed{23.2 \pm 0.65} \Rightarrow (22.55, 23.85)$

so "90% confident" that μ is in interval: $(22.55, 23.85)$

Now, continuing with above data ($n = 400$, $\bar{x} = 23.2$, $S = 8.0$)

So, if wanted to do a "two-sided" test of hypothesis

null hypothesis: $H_0: \mu = 24.0$ vs.

Alternative hypothesis: $H_A: \mu \neq 24.0$ AT $\boxed{\alpha = .1}$ LEVEL,

WE CONSTRUCT THE TEST STATISTIC: $\boxed{\frac{\bar{X} - 24.0}{S / \sqrt{n}}} \underset{\substack{H_0 \\ \mu = 24.0}}{\sim} Z.$

WE'LL REJECT H_0 IF OUR DATA'S VALUE OF TEST STATISTIC IS "EXTREME" VS WHAT WOULD BE SUGGESTED BY H_0.

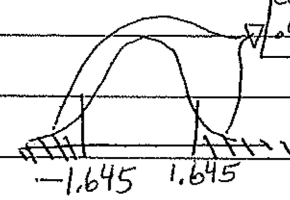
DATA'S VALUE OF TEST STATISTIC IS:

$$\frac{23.2 - 24.0}{\frac{8.0}{\sqrt{400}}} = \boxed{2.0}$$

Now, A (2-SIDED) $\alpha = .1$ REJECTION REGION

FOR A STD. NORMAL (Z) IS:

Combined area $.05 + .05 = .1$



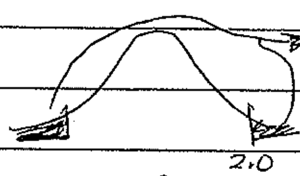
A VALUE OF TEST STATISTIC
THAT'S GREATER THAN 1.645 OR LESS THAN −1.645.

HERE OUR VALUE $\boxed{2.0}$ IS IN THE REJECTION REG.,
SO WE REJECT H_0 AND CLAIM H_A. $(\mu \neq 24.0)$

AN EQUIVALENT WAY TO CONDUCT THIS TEST

IS WITH P-VALUES (OBSERVED SIGNIFICANCE LEVEL).

HERE PVALUE =



$.026 + .026 = .052$ which is less than critical set α level of .10, so reject H_0 & claim H_A

DEFINITION:

PVALUE $\equiv P_{H_0}$ (GET A VALUE OF TEST STATISTIC THAT IS AT LEAST AS SUPPORTIVE OF H_A AS THE DATA'S VALUE OF TEST STATISTIC)

FOR THIS DATA,

NOTE: THE CONFIDENCE INTERVAL (C.I) RESULT AND THE (2-sided) TEST OF HYPOTHESIS (T.H.) RESULT ARE COMPLIMENTARY, IN SENSE THAT THE 90% i.e. $(1-\alpha)*100\%$ C.I FOR μ DOES NOT CONTAIN 24.0, WHILE THE $\alpha=.1$ T.H. THAT $\boxed{H_0: \mu=24.0}$ IS REJECTED AND IT IS CLAIMED THAT $\mu \neq 24.0$

n LARGE: TEST ON POPULATION PROPORTION: p

RECALL $\hat{p}$ (SAMPLE PROPORTION) IS LIKE A SAMPLE AVG OF 0's AND 1's; $SD(\hat{p})=\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ $E(\hat{p})=p$, so. FOR

$H_0: p=p_0$ $\quad \boxed{\dfrac{\hat{p}-p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}}} \underset{\substack{p=p_0 \\ n \text{ large}}}{\sim} Z$

so e.g. toy test ($\alpha=.1$ level) $H_0: p=.5$, $H_A: p>.5$, one-sided here

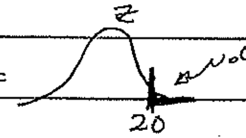
would reject H_0 if data value of test stat is "sufficiently large".

Ex. $n=100$ flips of coin by Tom; he gets \#1 for each head; we get \#1 for each tail. In the 100 flips, 60 heads (& thus 40 tails) were obtained.

H_0: "coin fair" ($p=.5$ get head)
H_A: coin ~~biased for heads~~ $p>.5$.

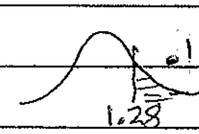
DATA'S VALUE OF TEST STAT:

$$\frac{\hat{p}-p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} = \frac{\frac{60}{100}-.5}{\frac{\sqrt{.5(1-.5)}}{\sqrt{100}}} = +2.0 \Rightarrow P_{VALUE} =$$



since P_{VALUE} OF .024 $< .1$, $\Rightarrow$ REJECT H_0, CLAIM H_A

ALTERNATIVELY $\alpha=.1$ (one-sided) REJECTION REGION IS:



DATA VALUE OF TEST STAT > 1.28, WHICH IT IS.

(TEST OF HYPOTHESIS)

n small T.H. ON $\boxed{\mu}$ UNDER NORMAL POPULATION:

RECALL HERE $\boxed{\dfrac{\bar{X}-\mu}{S/\sqrt{n}}} \sim t$ WITH n-1 d.f.

$t_{n-1\,df}$, α, 0, t_α

SO, e.g. to test $\boxed{\begin{array}{l} H_0: \mu = \mu_0 \\ H_A: \mu > \mu_0 \end{array}}$ THEN REJECT H_0 IFF "OBSERVED" $\boxed{\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}} > t_\alpha^{df=n-1}$

(α critical level)

if instead test $\boxed{\begin{array}{l} H_0: \mu = \mu_0 \\ H_A: \mu < \mu_0 \end{array}}$ THEN REJECT H_0 IFF $\boxed{\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}} < -t_\alpha$

$t_{n-1\,df}$, α, $-t_\alpha$

To test $\boxed{\begin{array}{l} H_0: \mu = \mu_0 \\ H_A: \mu \neq \mu_0 \end{array}}$ THEN REJECT H_0 IFF $\boxed{\left(\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{\alpha/2}\right) \text{ OR } \left(\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} < -t_{\alpha/2}\right)}$

2-sided test

$\alpha/2$, $\alpha/2$, $-t_{\alpha/2}$, $t_{\alpha/2}$

THE 2-sided T.H. result is complimentary to the corresponding confidence interval (C.I) result of $\boxed{\bar{x} \pm t_{\alpha/2}^{n-1\,df} \frac{s}{\sqrt{n}}}$ IN SENSE THAT IF μ_0 IS NOT IN C.I., THEN WOULD REJECT $H_0: \mu = \mu_0$

ACTUAL DATA EX.: SAY THE "ALLOWED" AVERAGE CONTAMINATION LEVEL IN A LAKE IS 30.0 $\frac{g}{cm^3}$, SAY TOOK n=16 SAMPLES TAKEN & GOT AVG RESULT OF 56.5, WITH SD OF RESULTS = 25.0, CAN ONE "PROVE" AT α=.05 LEVEL, THAT LAKE CONTAMINATION LEVEL IS ABOVE ALLOWABLE (EXCEEDS)

$\boxed{\begin{array}{l} \text{TEST } H_0: \mu = 30.0 \\ H_A: \mu > 30.0 \end{array}}$ Test Stat: $\dfrac{\bar{X}-30.0}{S/\sqrt{n}} \underset{H_0 \,\&\, \text{NORMAL INDIV. LEVELS}}{\sim} t$ with $\boxed{16-1}$ df = 15

WILL REJECT H_0, claim H_A & THUS "PROVE" (@ α=.05 LEVEL) EXCESS:

IF $\boxed{(\text{OBSERVED}) \dfrac{\bar{x}-30.0}{s/\sqrt{n}} > t_{.05}^{df=15}}$ here $\dfrac{\bar{x}-30.0}{s/\sqrt{n}} = \dfrac{56.5-30}{25/\sqrt{16}} \doteq 4.24 > t_{.05}^{df=15}$

$\Rightarrow$ PROVED EXCESS

TEST OF HYPOTHESIS ON $\boxed{\sigma^2}$ UNDER NORMAL POPULATION

RECALL IF $X_1, X_2, \ldots X_n$ RANDOM (INDPT) SAMPLE FROM NORMAL(μ, σ).

THEN $\boxed{\dfrac{(n-1)S^2}{\sigma^2}} \sim \chi^2$ WITH $n-1$ d.f.

SO, FOR NORMAL POPULATION, CAN USE THIS TO PERFORM A TEST ON σ^2 (POP'L VARIANCE)

Ex. GO BACK TO OUR LAKE CONTAMINATION STUDY, WHERE "ASSUMED" THAT INDIV. CONTAMINATION LEVELS IN LAKE ARE NORMALLY DISTRIBUTED.

BUT NOW WANT TO DO A TEST ON (POPUL) VARIANCE σ^2 OF INDIV. LEVELS IN LAKE (POPUL)

RECALL SAMPLE SIZE: $n = 16$.

RECALL SAMPLE SD, $S = 25.0$, SO, SAMPLE VARIANCE $S^2 = 25^2 = 625$

SAY WANT TO DO $\alpha = .05$ LEVEL TEST: $H_0: \sigma^2 = 500$ (i.e. $\sigma = \sqrt{500}$)

vs $H_A: \sigma^2 > 500$ (i.e. $\sigma > \sqrt{500}$)

THEN TEST STATISTIC IS: $\dfrac{(16-1)(S^2)}{500} \underset{H_0 \,\&\, \text{NORMAL INDIV LEVELS}}{\sim} \chi^2$ WITH $\underbrace{16-1}_{15}$ d.f.

THEN REJECT H_0 & CLAIM H_A IFF

$$\frac{(16-1)\cdot 625}{500} > \chi^{2\;\;15\text{ d.f.}}_{.05}$$

INTERPRETATION:

"SIGN TEST"
(Chapter 14: DISK SUPPLEMENTAL)

n small | ~~DO~~ CAN NOT ASSUME NORMAL POPULATION.

YOU CAN STILL DO A TEST ON

THE POPULATION MEDIAN, η, AS LONG

AS ANY CONTINUOUS POPULATION.

RECALL FOR ANY CONTINUOUS POPUL., $P(X_i < \eta) = .5 = P(X_i > \eta)$

FOR EX., LET'S LOOK AT SIMPLER VERSION OF "CONTAMINATED"

LAKE, WHERE ONLY $n = 7$ SAMPLES ARE TAKEN,

WE DO NOT ASSUME NORMAL (BUT ONLY CONTINUOUS ✓)

DISTRIB. OF LEVELS IN LAKE; SAY NOW CAN TEST (@ $\alpha = .05$ LEVEL)

POP'L MEDIAN

$H_0: \eta = 30.0$

$H_A: \eta > 30.0$

& SUPPOSE OUR $n=7$ LEVELS ARE:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
35.8	24.2	33.4	45.8	39.2	36.1	38.7

NOTE THAT 6 OF THE 7 VALUES EXCEED 30.0 (ALL BUT ~~THE 24.2~~)

HERE,

TEST STATISTIC IS: $\boxed{\text{\# OF } X_i\text{'S THAT EXCEED 30.0}} \underset{\substack{H_0 \\ \eta = 30.0}}{\sim} \text{BINOMIAL}(n=7, p=0.5)$

$\text{PVALUE} = P(\text{Binomial}(n=7, p=0.5) \geq 6) = \boxed{1 - P(\text{BIN}(7, .5) \leq 5)} < 0.05$

OBSERVED VALUE TEST STAT

CAN USE BINOMIAL TABLE

critical α

SO, REJECT H_0 &

SO, CLAIM $\eta > 30.0$

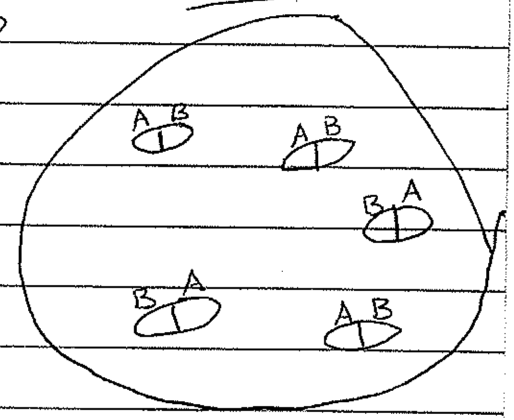
TWO POPULATIONS: PAIRED DATA

EXAMPLE: SUPPOSE $n=5$ PLOTS ARE RANDOMLY SELECTED FROM A REGION. EACH PLOT IS "SPLIT" IN TWO, WITH ONE HALF RANDOMLY SELECTED TO GET TRT A, WHILE THE OTHER HALF GETS TRT B. THEN @ END OF GROWING SEASON, THE CORN YIELD IS MEASURED ON EACH "HALF PLOT".

SUPPOSE RESULTING DATA IS:

(PAIR) PLOT	Trt A yield	Trt B yield	A-B yield difference
1	24	23	+1
2	32	33	−1
3	37	37	0
4	36	34	2
5	40	37	3

REGION



SO, IN ANALYSIS, VIEW THIS AS A SAMPLE OF $n=5$ DIFFERENCES OUT OF A "SINGLE" POPULATION OF ALL POSSIBLE PAIRED (A-B) DIFFERENCES.

SO, E.G. IF ASSUME POP'L OF DIFFERENCES IS ABOUT NORMAL, THEN CAN DO A t-test ON μ_{DIFF} (THE POP'L MEAN DIFF)

Trts A & B equally effective

$H_0: \boxed{\mu_{DIFF}=0}$

$H_A: \mu_{DIFF} \neq 0$

Test stat: $\dfrac{\bar{X}_{diff}-0}{S_{diff}/\sqrt{5}} \sim t$ with $5-1=4$ df.

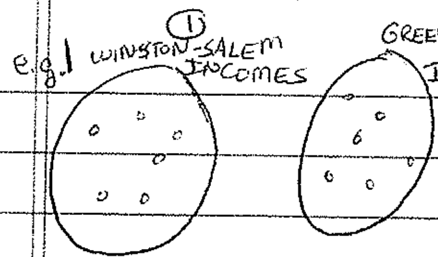
NOTE: $\bar{X}_{diff} = \boxed{\dfrac{+1+(-1)+0+2+3}{5} = +1}$ & $\boxed{s = \sqrt{\dfrac{\Sigma(d_i-\bar{X}_{diff})^2}{n-1}} = 1.58}$ $\Rightarrow$ obs test stat $= 1.41 \not> t_{.05}$ (4 df),

$\equiv \boxed{\bar{d}}$

so do NOT rej. H_0

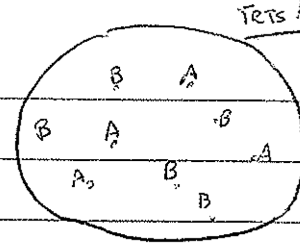
TWO INDEPENDENTLY SAMPED POPULATIONS.

e.g.1 ① WINSTON-SALEM INCOMES ② GREENSBORO INCOMES



OR

e.g.2 TETS A, B



● TESTS ON DIFFERENCES IN POPULATION MEANS μ_1, μ_2.

$H_0: \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$

● LARGE SAMPLES SETTING:

IF n_1 AND n_2 LARGE, THEN BY CLT,

☆ $$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{SD(\bar{X}_1 - \bar{X}_2)}} \sim Z$$

$\sigma_{(\bar{X}_1 - \bar{X}_2)}$

NOTE: $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \boxed{\mu_1 - \mu_2}$

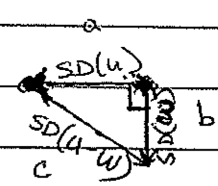
RECALL $SD(\bar{X}_i) = \sigma_{\bar{X}_i} = \frac{\sigma_i}{\sqrt{n_i}}$

GENERAL RULE: IF 2 THINGS U & W ARE INDPT,

(PYTHAGOREAN THEOREM) THEN $SD(U-W) = \sqrt{[SD(U)]^2 + [SD(W)]^2}$

$c^2 = a^2 + b^2$

SO $c = \sqrt{a^2 + b^2}$



SO ☆ ABOVE BECOMES $$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}}} \sim Z$$

Recall n_i large that $s_i \approx \sigma_i$

2 INDPTLY SAMPLED POPULATIONS; n_i's LARGE (CONTINUED)

so to test $H_0: (\mu_1 - \mu_2) = 0$

use Test statistic

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\left(\frac{S_1}{\sqrt{n_1}}\right)^2 + \left(\frac{S_2}{\sqrt{n_2}}\right)^2}}}$$ which is $\underset{H_0}{\sim} Z$

For $H_A: (\mu_1 - \mu_2) \neq 0$ [i.e. for $H_A: \mu_1 \neq \mu_2$], Rej H_0 & Claim H_A

iff observed value of test statistic is $\begin{cases} > Z_{\alpha/2} \\ \text{OR} \\ < -Z_{\alpha/2} \end{cases}$

NOTE: (FOR LARGE n_1, n_2), THE ABOVE TEST CAN BE "INVERTED" TO OBTAIN THE $(1-\alpha)*100\%$ C.I FOR $(\mu_1 - \mu_2)$

IT IS $$\boxed{(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \cdot \sqrt{\left(\frac{S_1}{\sqrt{n_1}}\right)^2 + \left(\frac{S_2}{\sqrt{n_2}}\right)^2}}$$

IT IS COMPLIMENTARY TO THE ABOVE TEST IN SENSE THAT ~~IF~~ [$(\mu_1 - \mu_2)$ VALUE OF] 0 is NOT in THE ABOVE C.I, THEN WOULD REJECT H_0

2 INDPTLY SAMPLED POPULATIONS.
n_1 AND n_2 SMALL

IF Assume both populations ARE NORMAL
& HAVE A COMMON VARIANCE

(i.e. $X_{11}, X_{12}, \ldots X_{1n_1} \sim \text{Normal}(\mu_1, \sigma^2)$ the same variance, but unknown.
$X_{21}, X_{22} \ldots X_{2n_2} \sim \text{Normal}(\mu_2, \sigma^2)$)

THEN to do A TEST ON $H_0: \boxed{\mu_1 - \mu_2 = 0}$

You first get a pooled estimate for σ^2,

namely $$\boxed{s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1-1+n_2-1}} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_1-1+n_2-1}$$

Now, test statistic looks like last one, except put in s_p for both s_1 & s_2

AND NOW (since n's small, s_p is NOT great est of σ^2)

the test statistic has a t distribution (with n_1-1+n_2-1 df)

Specifically $$\frac{\bar{X}_1 - \bar{X}_2 - (0)}{\sqrt{\left(\frac{s_p}{\sqrt{n_1}}\right)^2 + \left(\frac{s_p}{\sqrt{n_2}}\right)^2}} \underset{H_0}{\sim} t \text{ with } n_1-1+n_2-1 \text{ d.f.}$$

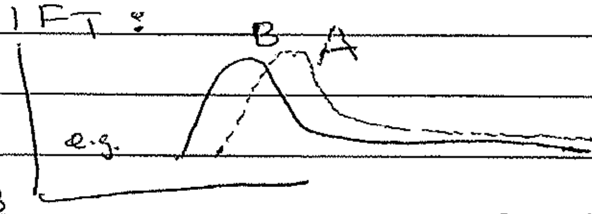
, so for $H_A: \mu_1 - \mu_2 \neq 0$, Rej H_0 iff test stat value $> t_{\alpha/2}$ OR $< -t_{\alpha/2}$

corresponding $(1-\alpha) * 100\%$ CI for $(\mu_1 - \mu_2)$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}^{df=n_1-1+n_2-1} \cdot \sqrt{\left(\frac{s_p}{\sqrt{n_1}}\right)^2 + \left(\frac{s_p}{\sqrt{n_2}}\right)^2}$$

2 INDPTLY SAMPLED POPULATIONS
n_1 AND n_2 SMALL

CAN/DO NOT NECESSARILY ASSUME NORMAL DISTRIBUTIONS, ONLY ASSUME CONTINUOUS DISTRIBUTIONS WITH POSSIBLE LOCATION SHIFT:



To Test H_0: A & B have same dist.

H_A: A is shifted to right of B
(i.e. $P(X_{A_j} > X_{B_k}) > \frac{1}{2}$) ← more likely that obs from A is larger than obs from B

Use "Sum-RANK Test" (Wilcoxon) - SUPPLEMENTAL DISK Chpt 14.

Ex. Emphasizing Concept

SAY GOT $n_A = 5$ VALUES FROM A OF: 17.1, 22.3, 9.2, 20.5, 15.9
& INDPTLY GOT $n_B = 4$ VALUES FROM B OF: 8.1, 7.4, 12.2, 6.9, 8.5

NOW IF RANK ALL 5+4=9 VALUES (IN ORDER) GET:

6.9, 7.4, 8.1, 8.5, 9.2, 12.2, 15.9, 17.1, 20.5, 22.3
(B B B B B B A A A A.)

NOTE! under H_0 OF A & B having SAME dist, that the 5 positions for the A's in the 9 total "ordered" slots are equally likely, there are $\binom{9}{5} = \frac{9!}{5!(9-5)!} = 126$ total (equally likely under H_0) sets of 5 positions for the A,

only 2 of these are at least as supportive of H_A as our observed positions (namely our observed & (BBBBAAAAA))
(BBBABAAAA) $\underset{4\quad 6 \longrightarrow 9}{}$ $\underset{5 \longrightarrow 9}{}$

so Pvalue for our test is $\frac{2}{\binom{9}{5}} = \frac{2}{126} \doteq \boxed{.016}$; so e.s. would reject H_0 @ $\alpha = .05$ level

TWO INDPT POPULATIONS

ASSUME BOTH POPULATIONS NORMAL

TEST ON POPULATION VARIANCES

SPECIFICALLY $X_{11}, X_{12} \cdots X_{1n_1} \overset{\text{indpt}}{\sim} \text{Normal}(\mu_1, \sigma_1^2)$

$X_{21}, X_{22} \cdots X_{2n_2} \overset{\text{indpt}}{\sim} \text{Normal}(\mu_2, \sigma_2^2)$

Recall we know: $\dfrac{(n_i-1)\, S_i^2}{\sigma_i^2} \sim \chi^2$ with d.f $n_i - 1$

For either pop'l i

$$* \quad \text{Now} \quad \frac{\overbrace{\dfrac{(n_1-1)\,S_1^2}{\sigma_1^2}}^{\sim \chi^2 \text{ w/ } n_1-1 \text{ df}} \Big/ (n_1-1)}{\underbrace{\dfrac{(n_2-1)\,S_2^2}{\sigma_2^2}}_{\sim \chi^2 \text{ w/ } n_2-1 \text{ df}} \Big/ (n_2-1)} \quad \boxed{\text{OF FORM}} \quad \text{indpt} \quad \left\{ \frac{\chi^2_{\text{"1"}} \big/ \text{df for } \chi^2_{\text{"1"}}}{\chi^2_{\text{"2"}} \big/ \text{df for } \chi^2_{\text{"2"}}} \right\} = F_{\substack{\text{numdf} = \text{df for "1"} \\ \text{dendf} = \text{df for "2"}}}$$

NOTE: THAT IF $\sigma_1^2 = \sigma_2^2$ $\left[\text{i.e. if } \dfrac{\sigma_1^2}{\sigma_2^2} = 1\right]$, then left-hand side $*$

simplifies to just $\boxed{\dfrac{S_1^2}{S_2^2}}$

so to test $H_0: \sigma_1^2 = \sigma_2^2$ $\left(\text{i.e. } \dfrac{\sigma_1^2}{\sigma_2^2} = 1\right)$ vs $H_A: \sigma_1^2 > \sigma_2^2$ $\left(\text{i.e. } \dfrac{\sigma_1^2}{\sigma_2^2} > 1\right)$

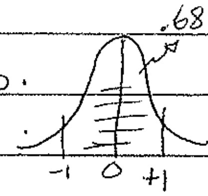
use test stat. of $\dfrac{S_1^2}{S_2^2} \underset{H_0}{\sim} F_{\substack{\text{numdf} = n_1 - 1 \\ \text{dendf} = n_2 - 1}}$ & rej H_0 iff $\boxed{\text{obs. } \dfrac{S_1^2}{S_2^2}} > F_{\alpha,\ \substack{\text{numdf} = n_1-1 \\ \text{dendf} = n_2-1}}$

$\boxed{\text{ex}}$ sample of size 16 from pop'l 1 got $s_1^2 = 6.2$
(indpt) " " " 21 " 2 got $s_2^2 = 2.1$

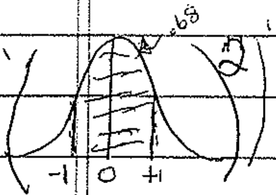
then obs $\dfrac{S_1^2}{S_2^2} = \dfrac{6.2}{2.1} = 2.95$ while $F_{.05,\ \substack{\text{numdf}=15 \\ \text{dendf}=20}} =$ ____ , so ____________

χ^2: $\boxed{\text{Chi-Square}}$ DISTRIBUTIONS

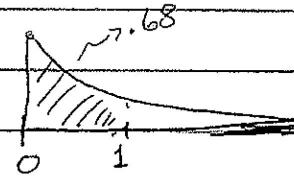
Start with a standard normal "Z" distrib.



Then Z^2 = "(



)^2" =



is a χ^2 with 1 d.f.

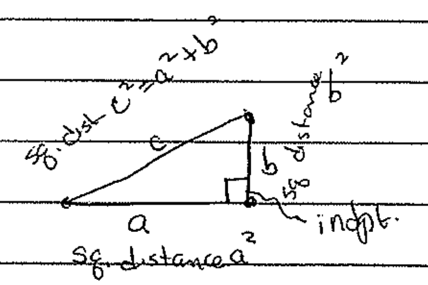
A χ^2 distrib. can also be thought of as a ("random" "distance")^2

GENERAL RESULT.

The sum of independent χ^2's is a χ^2

whose d.f. is the sum of the individual χ^2's d.f.'s.

This is an "extension of Pythagorean Thm"



$\boxed{\text{FOR } X_1, X_2, \ldots X_n \sim \text{Normal}(\mu, \sigma)}$

NOTE: $\boxed{\text{"t" distrib with } n-1 \text{ df}} = \boxed{\dfrac{\bar{X}-\mu}{s/\sqrt{n}}} = \boxed{\dfrac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{s}{\sigma}}} = \boxed{\dfrac{Z}{\frac{s}{\sigma}}}$

$$\boxed{\dfrac{Z}{\sqrt{\dfrac{\chi^2(n-1\,df)}{n-1}}}} \leftarrow \boxed{\dfrac{Z}{\sqrt{\dfrac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}}}$$

If $X_i \sim$ Normal(μ, σ) & indpt : Then Note: $\sum_{i=1}^{n} \left(\dfrac{X_i - \mu}{\sigma}\right)^2 \equiv \sum_{i=1}^{n} (Z_i^2) = \boxed{\text{sum of } n \text{ indpt } \chi^2(1\text{'s})} \Rightarrow \chi^2(n \text{ d.f.})$.

Now, $\boxed{\sum_{i=1}^{n} \left(\dfrac{X_i - \bar{X}}{\sigma}\right)^2} = \sum_{i=1}^{n} \left(\dfrac{X_i - \mu}{\sigma}\right)^2 - \left(\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \chi^2(n) - \chi^2(1) = \chi^2(n-1 \text{ d.f.})$.

$\boxed{\dfrac{(n-1)s^2}{\sigma^2}} \sim \chi^2(n-1)$ df.

MANY PROBABILISTIC MODELS ARE OF FORM:

$$\boxed{Y_i = \underbrace{E(y_i)}_{\text{deterministic part}} + \underbrace{\epsilon_i}_{\text{random part.}}}$$

FOR person $i = 1, 2, \dots n$

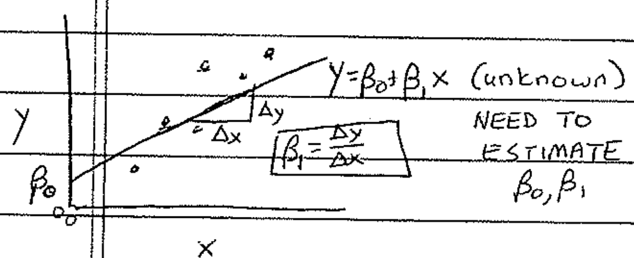
e.g. Simple Linear Regression of Weight of Person i (y_i) WITH HEIGHT of Person i (X_i)

is: $Y_i = \underbrace{\beta_0 + \beta_1 \cdot X_i}_{\text{simple linear deterministic part}} + \epsilon_i$ where $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, ~~$\epsilon_i$ are normal~~ $\boxed{\epsilon_i \sim normal(0, \sigma)}$

AND ϵ_i's are indpt.

here, β_0, β_1 and σ^2 are (unknown) parameters

Observe n (x_i, y_i) pairs:



$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$

choose $\hat{\beta}_0, \hat{\beta}_1$, so as to minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

choose $\hat{\beta}_0, \hat{\beta}_1$ to minimize $\boxed{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i))^2} \equiv f(\hat{\beta}_0, \hat{\beta}_1)$

Take $\frac{\partial f}{\partial \beta_0}$ set $= 0$

Take $\frac{\partial f}{\partial \beta_1}$ set $= 0$

to get $\hat{\beta}_0, \hat{\beta}_1$

Estimate σ^2 by: $\boxed{\frac{\sum (y_i - \hat{y}_i)^2}{n-1-1}}$ (est β_0, est β_1)

now label this as s^2

Now, test if x ~~(here height)~~ has a significant (linear) "effect" on y (here weight) by:

H_0: $\beta_1 = 0$ (no effect of x on y

H_1: $\beta_1 \neq 0$.

Test statistic: $\boxed{\frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}}} \underset{H_0}{\sim} t$ with n-2 d.f.

Rej H_0 if data value test stat $> t_{\alpha/2}^{n-2\,df.}$ OR $< -t_{\alpha/2}$

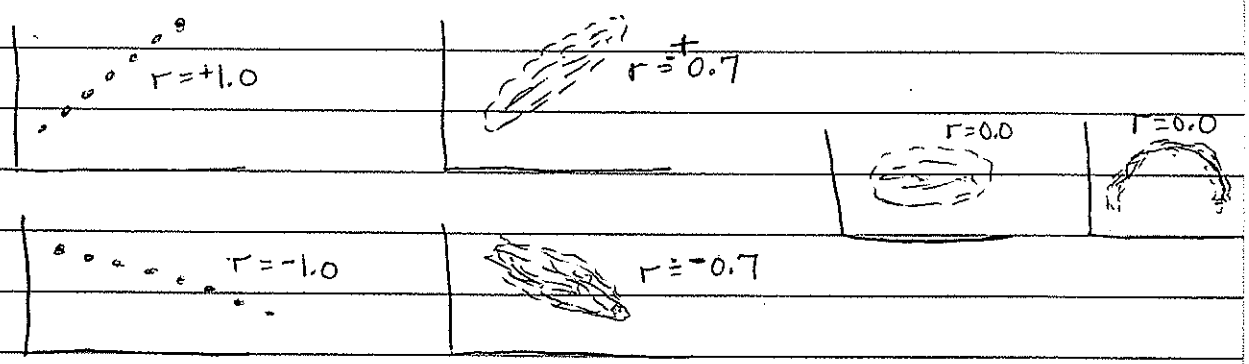
SAMPLE CORRELATION (r) BETWEEN X (here height) and y (here weight)

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\Sigma(x_i-\bar{x})^2\right]\left[\Sigma(y_i-\bar{y})^2\right]}} \quad \text{OR in book's notation} \quad = \boxed{\frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}}}$$

r can range from -1.0 to 0.0 to $+1.0$

- -1.0: perfect negative linear (x any y) association (for sample)
- 0.0: no net positive nor negative linear assoc.
- $+1.0$: perfect positive linear association

Some pictures:



long-run avg value for Y_i for given value of X_i

CONFIDENCE INTERVALS FOR $\boxed{E(Y_i)}$ FOR A GIVEN VALUE OF X_i:

$\boxed{\hat{\beta}_0+\hat{\beta}_1 X_i}$ $\boxed{\beta_0+\beta_1 X_i}$

FOR GIVEN X_i

$$\boxed{\frac{\hat{Y}_i - E(Y_i)}{\hat{\sigma}_{\hat{Y}_i}}} \sim t \text{ with } n-2 \text{ d.f.}$$

$\Rightarrow (1-\alpha)*100\%$ CI FOR $E(Y_i)$ is: $\boxed{\hat{Y}_i \pm t^{df=n-2}_{\alpha/2}\left(\hat{\sigma}_{\hat{Y}_i}\right)}$

PREDICTION INTERVALS FOR A FUTURE VALUE OF Y, SAY Y_F, FOR A GIVEN VALUE OF X: X_F

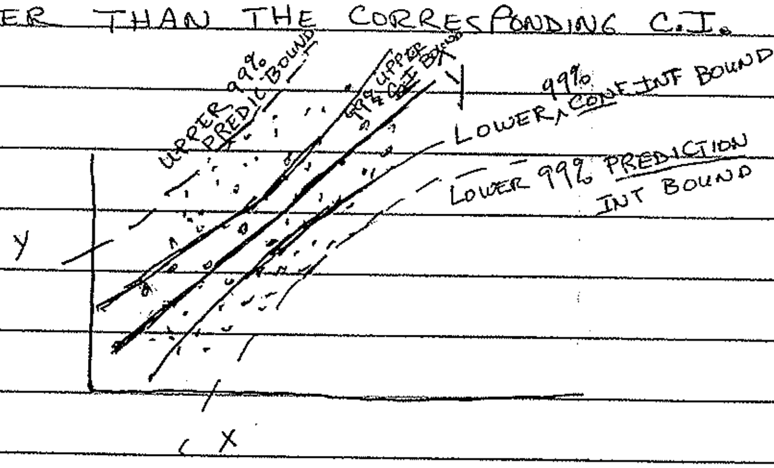
$$\boxed{\frac{Y_{F\,(@X_i)} - \hat{Y}_{\text{BASED ON } X_F}}{\hat{\sigma}_{(Y_F - \hat{Y}_{\text{BASED ON } X_F})}}} \sim t \text{ WITH } n-2 \text{ d.f.}$$

$\boxed{\text{OBTAINED USING PYTHAG. Thm}}$

$$\Rightarrow \hat{Y}_i \pm t_{\alpha/2}\left(\hat{\sigma}_{(Y_F - \hat{Y}_{\text{BASED ON } X_i})}\right)$$

THE PREDICTION INTERVAL WILL ALWAYS BE WIDER THAN THE CORRESPONDING C.I.

"ROUGHLY" GET RESULTS LIKE THE FOLLOWING:



MULTIPLE REGRESSION

NOW INSTEAD OF A SINGLE PREDICTOR (X)

WE HAVE MULTIPLE "TERMS", CALL THEM $Z_1, Z_2, \ldots, Z_K$

THAT CAN INFLUENCE Y.

THE MODEL IS NOW:

(FOR $i=1,2,\ldots,n$)

$$Y_i = \underbrace{\beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \cdots + \beta_k z_{ki}}_{\substack{E(Y_i) \\ \text{deterministic part}}} + \underbrace{\epsilon_i}_{\substack{\text{random} \\ \text{part.}}}$$

$\epsilon_i \sim \text{normal}(0, \sigma^2)$

ϵ_i's indpt.

(unknown) parameters are $\beta_0, \beta_1, \ldots, \beta_k$ & σ^2 which must be estimated

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 z_{1i} + \hat{\beta}_2 z_{2i} + \cdots + \hat{\beta}_k z_{ki}$$

Choose the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \ldots \hat{\beta}_k$ so as to minimize: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

i.e. Choose ests. so as to minimize $\sum_{i=1}^{n}\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 z_{1i} + \hat{\beta}_2 z_{2i} + \cdots + \hat{\beta}_k z_{ki})\right)^2$

conceivably could take partial derivative with respect to each β_j & set all $=0$, we'll see later that doing this with vector/matrix notation greatly simplifies the concepts/math.

Then σ^2 is estimated by $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(k+1)} \equiv S^2$

To test $H_0: \beta_j = 0$
$H_A: \beta_j \neq 0$

Use Test Stat: $\boxed{\dfrac{\hat{\beta}_j - 0}{\hat{\sigma}_{\hat{\beta}_j}}} \underset{H_0}{\sim} t$ with $n-(k+1)$ d.f.

VECTOR/MATRIX EXPRESSION OF MULTIPLE REGR.

- REVIEW SIMPLE LINEAR ALGEBRA ARGUMENTS
- REVIEW MATRIX LOOK OF MULTIPLE REGRESSION

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 \cdot 1 + \beta_1 z_{11} + \beta_2 z_{21} + \cdots + \beta_k z_{k1} + \epsilon_1 \\ \beta_0 \cdot 1 + \beta_1 z_{12} + \beta_2 z_{22} + \cdots + \beta_k z_{k2} + \epsilon_2 \\ \vdots \\ \beta_0 \cdot 1 + \beta_1 z_{1n} + \beta_2 z_{2n} + \cdots + \beta_k z_{kn} + \epsilon_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & z_{11} & z_{21} & \cdots & z_{k1} \\ 1 & z_{12} & z_{22} & & z_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{1n} & z_{2n} & & z_{kn} \end{pmatrix}}_{\text{label as "Z"}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\text{label as } \underline{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\text{label as } \underline{\epsilon}}$$

$\mathcal{Z}$ NOT NORMAL, IT'S A MATRIX OF "PREDICTORS"

$$\Rightarrow \boxed{\underline{y} = Z \cdot \underline{\beta} + \epsilon}$$

$$\Rightarrow \hat{\underline{y}} = Z\hat{\underline{\beta}} = Z \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

in vector/matrix notation

Choose $\hat{\beta}$ so as to minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\underline{y} - \hat{\underline{y}})^T (\underline{y} - \hat{\underline{y}}) = (\underline{y} - Z\hat{\underline{\beta}})^T (\underline{y} - Z\hat{\underline{\beta}})$

$$= (\underline{y}^T - \hat{\underline{\beta}}^T Z^T)(\underline{y} - Z\hat{\underline{\beta}}) = \underline{y}^T\underline{y} \underbrace{- \hat{\underline{\beta}}^T Z^T \underline{y} - \underline{y}^T Z \hat{\underline{\beta}}}_{-2\hat{\underline{\beta}}^T Z^T \underline{y}} + \hat{\underline{\beta}}^T Z^T Z \hat{\underline{\beta}}$$

Now Take $\frac{\partial^{k+1}}{\partial \underline{\beta}} \Rightarrow 0 - 2Z^T\underline{y} + 2(Z^TZ)\hat{\underline{\beta}}$ set $= 0 \Rightarrow \boxed{\hat{\underline{\beta}} = (Z^TZ)^{-1} Z^T \underline{y}}$

NOTE: $E(\hat{\underline{\beta}}) = E\left[(Z^TZ)^{-1} Z^T \underline{y}\right] = (Z^TZ)^{-1} Z^T E(\underline{y}) = (Z^TZ)^{-1} Z^T (Z\underline{\beta}) = (Z^TZ)^{-1}(Z^TZ)\underline{\beta}$

Using $E(cX) = cE(X)$

$= \underline{\beta}$ so $\hat{\underline{\beta}}$ is unbiased est. of $\underline{\beta}$

$$\boxed{\begin{matrix}\text{Variance} \\ \text{Cov MATRIX} \\ \text{FOR } \hat{\beta}\end{matrix}} = (Z^TZ)^{-1} Z^T \begin{pmatrix} \text{Var} \\ \text{Cov} \\ \text{FOR } \underline{y} \end{pmatrix} \left[(Z^TZ)^{-1} Z^T\right]^T = (Z^TZ)^{-1} \sigma^2$$

Using $Var(c \cdot X) = c \cdot Var(X) \cdot c^T$

$$\begin{pmatrix} Var\,\hat{\beta}_0 & Cov(\hat{\beta}_0, \hat{\beta}_1) & & \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var\,\hat{\beta}_1 & \boxed{Cov(\hat{\beta}_m, \hat{\beta}_k)} & \\ & & Var\,\hat{\beta}_j & \\ & & & Var(\hat{\beta}_k) \end{pmatrix}$$

generic

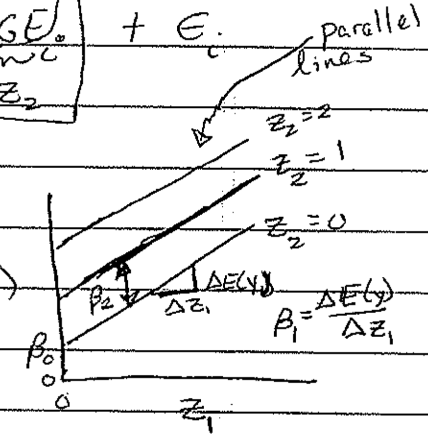
so $\widehat{Var}\,\hat{\beta}_j = \left[(j,j) \text{ element of } (Z^TZ)^{-1}\right] \cdot \hat{\sigma}^2$

use this in t-test on β_j.

FIRST-ORDER MULTIPLE REGRESSION

- ALL OF THE PREDICTING "TERMS" $(z_1, z_2 \cdots z_k)$ ARE SEPARATE LINEAR (1^{ST}-ORDER) VARIABLES

Ex. $\underbrace{\boxed{\text{WEIGHT}}_i}_{y} = \underbrace{\boxed{\beta_0 + \beta_1 \cdot \underbrace{\boxed{\text{HEIGHT}}_i}_{z_1} + \beta_2 \cdot \underbrace{\boxed{\text{AGE}}_i}_{z_2}}}_{E(Y_i)} + \epsilon_i$



$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 z_{1i} + \hat{\beta}_2 z_{2i}}$, $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

$\underset{\text{matrix}}{Z} = \begin{pmatrix} 1 & z_{11} & z_{21} \\ 1 & z_{12} & z_{22} \\ \vdots & \vdots & \vdots \\ 1 & z_{1n} & z_{2n} \end{pmatrix}$ $\quad \hat{\underline{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (Z^T Z)^{-1} Z^T \underline{y}$

$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \hat{\underline{y}} = Z \hat{\underline{\beta}}$; $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-3} = \frac{(\underline{y} - \hat{\underline{y}})^T (\underline{y} - \hat{\underline{y}})}{n-3}$

To test $H_0: \beta_j = 0$ vs $H_A: \beta_j \neq 0$.

Test stat: $\boxed{\frac{\hat{\beta}_j - 0}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}} \underset{H_0}{\sim} t$ with $n-3$ df

(where $\widehat{Var}(\hat{\beta}_j) = [(j,j) \text{ element } (Z^T Z)^{-1}] \cdot \hat{\sigma}^2$)

Rej Ho iff $\boxed{\text{observed value of test stat}}$ $> t_{\alpha/2}$ OR $< -t_{\alpha/2}$

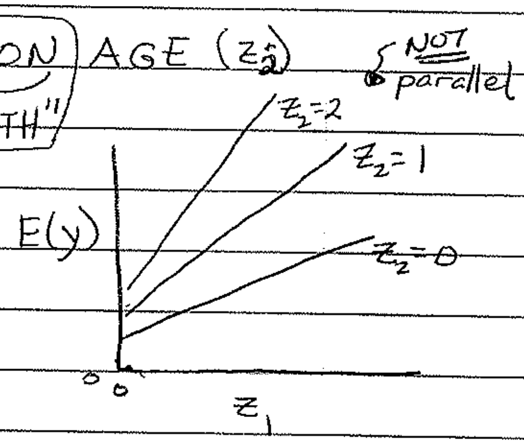
ALLOWING INTERACTONS BETWEEN VARIABLES

EX.

EXPAND THE $\underbrace{\text{WEIGHT}}_{y}$: $\underbrace{\text{HEIGHT}}_{z_1}$, $\underbrace{\text{AGE}}_{z_2}$ 1^{ST} ORDER

MODEL TO ALLOW THE LINEAR EFFECT

OF HEIGHT (z_1) TO $\boxed{\underbrace{\text{DEPEND ON}}_{\text{"INTERACT WITH"}}}$ AGE (z_2)



FOR $i=1,2,\ldots,n$,

$$\boxed{\text{WEIGHT}}_i = \underbrace{\beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 \underbrace{\boxed{z_{1i} \cdot z_{2i}}}_{z_{3i}}}_{E(y_i)} + \epsilon_i$$

Now the Z matrix has an extra 4^{th} column & $\underline{\beta}$ has 4^{th} value

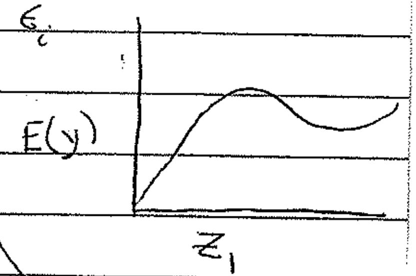
Still $\hat{\underline{\beta}}$ of form $(Z^T Z)^{-1} Z^T \underline{y}$ $\quad \hat{\underline{y}} = Z\hat{\beta}$ $\quad$ Now $\hat{\sigma}^2 = \dfrac{(\underline{y}-\hat{\underline{y}})^T(\underline{y}-\hat{\underline{y}})}{n-4}$

& NOTE can test if there is an interaction effect by testing $H_0: \beta_3 = 0$ in a similar t-test manner.

ALLOWING QUADRATIC (2^{ND}-ORDER) AND HIGHER ORDER TERMS

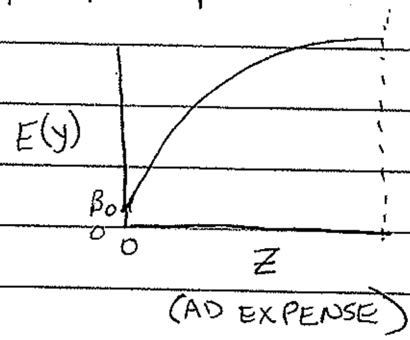
EX: EXPAND THE $\underbrace{\text{WEIGHT}}_{Y}$ & $\underbrace{\text{HEIGHT}}_{Z_1}$ $\underbrace{1^{ST}\text{ ORDER}}_{\text{LINEAR}}$ MODEL TO ALLOW QUADRATIC (2^{ND}-ORDER) & HIGHER ORDERS.

SAY $Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{1i}^2 + \beta_3 z_{1i}^3 + \epsilon_i$



KOHL'S SALES VS. ADVERTISING EXPENSE: Z
months $i=1,2,\ldots$

EX: $\underbrace{SALES_i}_{Y_i} = \beta_0 + \beta_1 z_i + \beta_2 z^2 + \epsilon_i$



HERE: β_2 is negative
β_1 is positive.

ALLOWING MULTIPLE VARIABLES OF NON-LINEAR ORDER AND INTERACTION(S) BETWEEN VARIABLES

BACK TO $\underbrace{\text{WEIGHT}}_{y}$: $\underbrace{\text{HEIGHT}}_{Z_1}$, $\underbrace{\text{AGE}}_{Z_2}$ BUT NOW ALLOW

2^{ND}-ORDER FOR HTGHT (Z_1) & AGE (Z_2), AND ALSO ALLOW INTERACTION BETWEEN HEIGHT & AGE @ LINEAR LEVEL,

LINEAR LEVEL INTERACTION

$$\Rightarrow Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{1i}^2 + \beta_3 Z_{2i} + \beta_4 Z_{2i}^2 + \beta_5 Z_{1i} \cdot Z_{2i} + \epsilon_i$$

(NOTE: IF ALSO ALLOWED INTERACTION BETWEEN Z_1 & Z_2 AT QUADRATIC (2^{ND}-ORDER) LEVEL,

THEN WOULD ALSO ADD A $\boxed{\beta_6 Z_{1i}^2 Z_{2i}^2}$ TERM.

INCORPORATING QUALITATIVE VARIABLES INTO MULTIPLE REGRESSION

BY USE OF "DUMMY" VARIABLES.

Ex. ONE WANTS TO COMPARE THE EFFECTS OF 3 DIFFERENT TEACHING (TUTORING) TECHNIQUES: $\boxed{A, B, C}$

ON STUDENT (TUTEE) PERFORMANCE (y).

• WE USE DUMMY "Z_j" VARIABLES SO WE CAN DO THIS IN A MULTIPLE REGRESSION SETTING. SET ONE LEVEL, SAY "A" AS BASE LEVEL.

THEN SET UP (0,1) DUMMY VARIABLES FOR EACH OTHER LEVEL.

HERE: $Z_{1i} = \begin{bmatrix} 1 \text{ if } i \text{ (TUTEE) GOT LEVEL } B \\ 0 \text{ if NOT} \end{bmatrix}$ $\quad Z_{2i} = \begin{bmatrix} 1 \text{ if } i \text{ GOT LEVEL } C \\ 0 \text{ if NOT} \end{bmatrix}$

TUTEE$_i$ PERFORMANCE

$$y_i = \underbrace{\boxed{\beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i}}}_{E(y_i)} + \epsilon_i$$

NOTE:
- if i got A, then $E(y_i) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$, so $\boxed{\mu_A = \beta_0}$
- if i got B, then $E(y_i) = \beta_0 + \beta_1(1) + \beta_2(0)$ so $\boxed{\mu_B = \beta_0 + \beta_1} \Rightarrow \boxed{\beta_1 = \mu_B - \mu_A}$
- if i got C, then $E(y_i) = \beta_0 + \beta_1(0) + \beta_2(1)$ so $\boxed{\mu_C = \beta_0 + \beta_2} \Rightarrow \boxed{\beta_2 = \mu_C - \mu_A}$

FOR DUMMY SETTING:

$\boxed{\beta_0 = \mu_{\text{BASE LEVEL}}}$; $\boxed{\beta_j = \mu_{\text{Level } Z_j} - \mu_{\text{BASE LEVEL}}}$

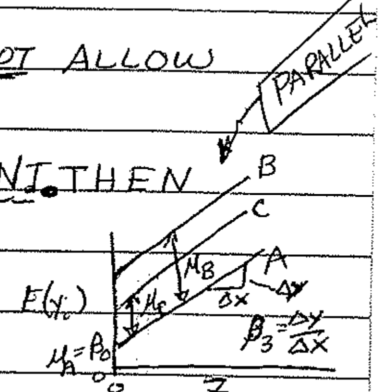
INCLUDING BOTH QUALITATIVE AND QUANTITATIVE VARIABLES IN MULTIPLE REGR.

QUALITATIVE: A, B, C

Ex. GO BACK TO "TUTEE/TECHNIQUE" EXAMPLE, BUT

QUANTITATIVE

NOW, ALSO ALLOW LINEAR EFFECT DUE TO LENGTH, z_{3i}, OF TUTEE i's TUTORING SESSION. 1ST DO NOT ALLOW INTERACTION BETWEEN QUALI & QUANT. THEN



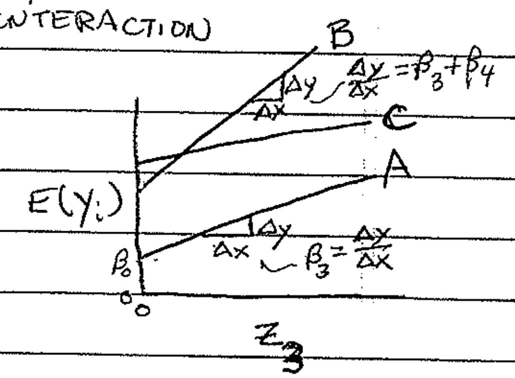
AGAIN $z_{1i} = \begin{bmatrix} 1 & \text{if } i \text{ got B} \\ 0 & \text{o.w.} \end{bmatrix}$ $\quad z_{2i} = \begin{bmatrix} 1 & \text{if } i \text{ got C} \\ 0 & \text{o.w.} \end{bmatrix}$

AND NOW MODEL IS: $Y_i = \beta_0 + \underbrace{\beta_1 z_{1i} + \beta_2 z_{2i}}_{\text{QUAL. Part}} + \underbrace{\beta_3 z_{3i}}_{\text{QUAN. part}} + \epsilon_i$

QL, QT INTERACTION

Ex. ABOVE BUT ALSO ALLOW INTERACTION BETWEEN QUAL & QUANT

Now, $Y_i = \beta_0 + \underbrace{\beta_1 z_{1i} + \beta_2 z_{2i}}_{\text{QUAL.}} + \underbrace{\beta_3 z_{3i}}_{\text{QUANT}} + \underbrace{\beta_4 z_{1i} z_{3i} + \beta_5 z_{2i} z_{3i}}_{\text{INTERACTION}} + \epsilon_i$



TEST OF HYPOTHESIS WHEN ONE MODEL (LABELLED "REDUCED") IS NESTED WITHIN A LARGER MODEL (LABELLED "COMPLETE")

GENERAL SETTING

VIEW "COMPLETE" (LARGER) MODEL AS:

$$\boxed{Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \cdots + \beta_g z_{gi} + \beta_{g+1} z_{g+1,i} + \cdots + \beta_K z_{Ki} + \epsilon_i}$$

AND "REDUCED" (SMALLER) MODEL AS:

$$\boxed{Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \cdots + \beta_g z_{gi} + \epsilon_i}$$

H_0: $0 = \beta_{g+1} = \beta_{g+2} = \cdots = \beta_K = 0$ (i.e. REDUCED MODEL "SUFFICES") (i.e. none of $z_{g+1}, z_{g+2}, \cdots, z_K$ are important in predicting Y)

vs H_A: Not H_0 (At least one of $z_{g+1}\ z_{g+2} \cdots z_K$ is important in predicting Y

1^{ST} (UNDER) FOR COMPLETE (C) MODEL, GET $\hat{\beta}_{0_c}, \hat{\beta}_{1_c}, \cdots \hat{\beta}_{K_c}$ to minimize $\sum (Y_i - \hat{Y}_{i_C})^2 \equiv SSE_C$

NEXT UNDER REDUCED (R) MODEL, GET $\hat{\beta}_{0_R}, \hat{\beta}_{1_R} \cdots \hat{\beta}_{g_R}$ to minimum $\sum (Y_i - \hat{Y}_{i_R})^2 \equiv SSE_R$

Test Statistic: $\boxed{\dfrac{\dfrac{SSE_R - SSE_C}{K-g}}{\dfrac{SSE_C}{n-(K+1)}}} \underset{H_0}{\sim} F_{\substack{numdf = K-g \\ dendf = n-(K+1)}}$

Rej H_0 iff observed value of test stat. $> F_\alpha$

ILLUSTRATION OF NESTED MODEL.

GO BACK TO THE $\boxed{\text{EX. QL, QT INTERACTION}}$

like z_{4i}^* like z_{5i}^*

$$* \quad \text{WHERE } Y_i = \beta_0 + \underbrace{\beta_1 z_{1i} + \beta_2 z_{2i}}_{\text{QUALI}} + \underbrace{\beta_3 z_{3i}}_{\text{QUANT}} + \underbrace{\beta_4 (z_{1i} z_{3i}) + \beta_5 (z_{2i} z_{3i})}_{\text{INTERACTION}} + \epsilon_i$$

NOW SUPPOSE ONE WANTED TO TEST THE NULL HYPOTHESIS

THAT NEITHER THE QUANT VARIABLE NOR INTERACTION ARE IMPORTANT IN PREDICTING y (i.e. $H_0: 0 = \beta_3 = \beta_4 = \beta_5 = 0$) VERSUS H_A: NOT H_0.

SO, COMPLETE MODEL IS GIVEN BY $*$, ABOVE, WITH $K=5$.

THE REDUCED MODEL IS: $y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \tilde{\epsilon}_i$, WITH $g=2$.

TEST STATISTIC:

$$\boxed{\frac{\frac{SSE_R - SSE_C}{K-g}}{\frac{SSE_C}{n-(K+1)}}} \underset{H_0}{\sim} \boxed{F_{\substack{\text{num d.f.} = K-g \\ \text{den d.f.} = n-(K+1)}}}$$

NOTE: $SSE_C = \Sigma(y_i - \hat{y}_{C_i})^2$; $SSE_R = \Sigma(y_i - \hat{y}_{R_i})^2$

Rej H_0 if observed value of test stat $> F_\alpha$ (for numdf $= K-g$, dendf $= n-(K+1)$)

VECTOR / MATRIX GEOMETRIC INTERPRETATION OF NESTED MODEL EX.

RECALL IN $\boxed{\text{QL, QT INTERACTION}}$ EX OF NESTED MODEL, THAT

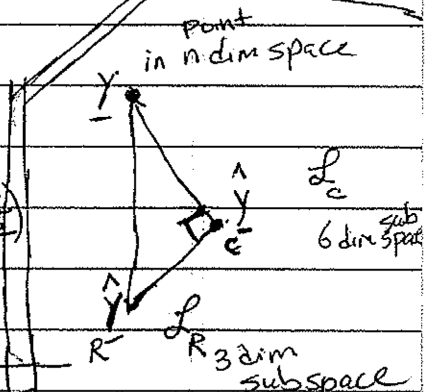
FOR $i=1,2,\dots,n$,

$$\boxed{{}_C\hat{Y}_i = {}_C\hat\beta_0\cdot 1 + {}_C\hat\beta_1 \cdot z_{1i} + {}_C\hat\beta_2 z_{2i} + {}_C\hat\beta_3 z_{3i} + {}_C\hat\beta_4 z_{4i} + {}_C\hat\beta_5 z_{5i}} \Rightarrow \begin{pmatrix} {}_C\hat{Y}_1 \\ {}_C\hat{Y}_2 \\ \vdots \\ {}_C\hat{Y}_n \end{pmatrix} \equiv {}_C\hat{\underline{y}} = \underbrace{\begin{pmatrix} 1 & z_{11} & z_{21} & z_{31} & z_{41} & z_{51} \\ 1 & z_{12} & z_{22} & z_{32} & z_{42} & z_{52} \\ 1 & \vdots & & & & \\ 1 & & & & & \\ 1 & z_{n1} & z_{n2} & z_{n3} & z_{n4} & z_{n5} \end{pmatrix}}_{Z} \begin{pmatrix} {}_C\hat\beta_0 \\ {}_C\hat\beta_1 \\ {}_C\hat\beta_2 \\ {}_C\hat\beta_3 \\ {}_C\hat\beta_4 \\ {}_C\hat\beta_5 \end{pmatrix}$$

$$= Z \cdot {}_C\hat{\underline\beta}$$

$$\Rightarrow {}_C\hat{\underline{y}} = \begin{pmatrix} {}_C\hat{Y}_1 \\ {}_C\hat{Y}_2 \\ \vdots \\ {}_C\hat{Y}_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{\underline{z}_0} {}_C\hat\beta_0 + \underbrace{\begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \\ z_{1n} \end{pmatrix}}_{\underline{z}_1} {}_C\hat\beta_1 + \underbrace{\begin{pmatrix} z_{21} \\ z_{22} \\ \vdots \\ z_{2n} \end{pmatrix}}_{\underline{z}_2} \hat\beta_2 + \underbrace{\begin{pmatrix} z_{31} \\ z_{32} \\ \vdots \\ z_{3n} \end{pmatrix}}_{\underline{z}_3} \hat\beta_3 + \underbrace{\begin{pmatrix} z_{41} \\ z_{42} \\ \vdots \\ z_{4n} \end{pmatrix}}_{\underline{z}_4} \hat\beta_4 + \underbrace{\begin{pmatrix} z_{51} \\ z_{52} \\ \vdots \\ z_{5n} \end{pmatrix}}_{\underline{z}_5} \hat\beta_5$$

with linear independence of $\underline{z}_0, \underline{z}_1, \dots \underline{z}_5$, ${}_C\hat{\underline{y}}$ is the point in the $5+1=6$ dimensional subspace of n dimensional space (generated by $\underline{z}_0, \underline{z}_1, \underline{z}_2, \underline{z}_3, \underline{z}_4, \underline{z}_5$) that has the smallest squared distance from $\underline{y}$



$$\boxed{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \equiv \underline{y}}$$

Similarly

$${}_R\hat{\underline{y}} = \begin{pmatrix} {}_R\hat{Y}_1 \\ {}_R\hat{Y}_2 \\ \vdots \\ {}_R\hat{Y}_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{\underline{z}_0} {}_R\hat\beta_0 + \underbrace{\begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \\ z_{1n} \end{pmatrix}}_{\underline{z}_1} {}_R\hat\beta_1 + \underbrace{\begin{pmatrix} z_{21} \\ z_{22} \\ \vdots \\ z_{2n} \end{pmatrix}}_{\underline{z}_2} {}_R\hat\beta_2$$

with linear indep. of $\underline{z}_0, \underline{z}_1, \underline{z}_2$, $\hat{\underline{y}}_R$ is the point in the $2+1=3$ dim (generated by $\underline{z}_0, \underline{z}_1, \underline{z}_2$) that has the smallest squared distance from $\underline{y}$

SSE_C is the squared distance between $\underline{y}$ and ${}_C\hat{\underline{y}}$

SSE_R " " " " $\underline{y}$ and ${}_R\hat{\underline{y}}$.