0282-1

1) p ~~XXV~~ statistics is the science of extracting information from data.

Ch1

1) p4 Individuals are the objects described by a data set. A variable is a characteristic recorded about an individual.

2) *p4 A categorical variable takes on several categories

eg race, hair color, gender

often count the number in each category or find the percentage adding the categories does not make sense

A quantitative variable takes on numerical values

eg height, # phone calls received adding the values makes sense

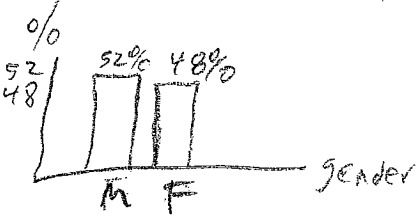
3) p4 The distribution of a variable tells what values it takes and how often.

4) p6 The distribution of a categorical variable lists counts (frequencies) or percents

ex SIU student gender M 52%
F 48%

5) p7 bar graph vertical axis height = percent or count

horiz axis has categories



A pie chart assigns percents of a circle area to categories
a,b) must use all of the categories

6) A (Frequency) histogram is a graph of the distribution of a quantitative variable. 1.5

7) p8-9 a) Divide the range of the data into classes of equal width. Each observation should fall in exactly one class.

ex semester credit hours 40 students

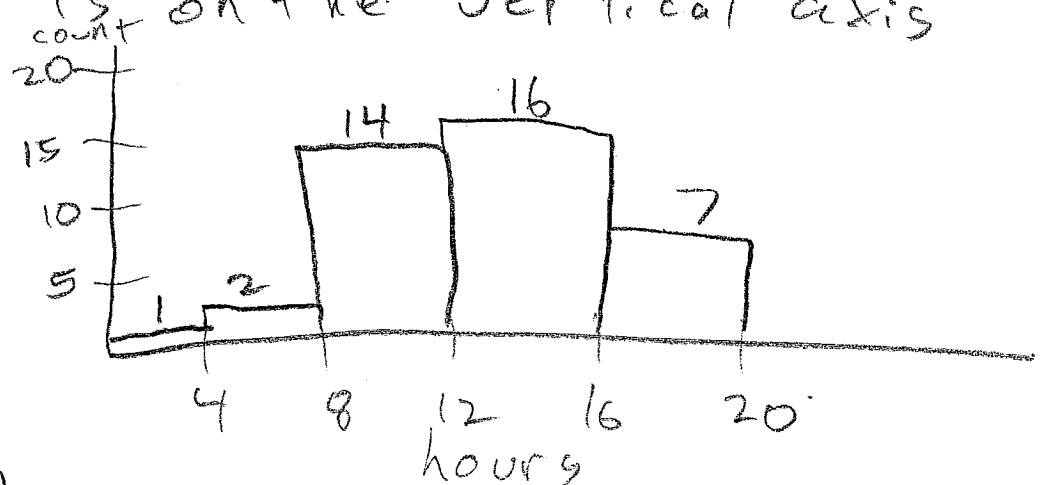
common mistake: 0-4, 4-8, 8-12, 12-16, 16-20

which class does 12 go into?

classes	count
$0 < \text{hours} \leq 4$	1
$4 < \leq 8$	2
$8 < \leq 12$	14 (12 goes in this class)
$12 < \leq 16$	16 (12 does not go in this class)
$16 < \leq 20$	7

b) find the count of observations with each class

c) On the horiz axis mark the scale of the variable. Bar height = count is on the vertical axis



d) label the top of the bar with the count (not in book)

between point 8 and 9 01:25

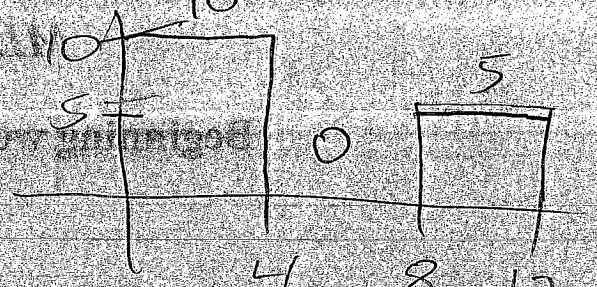
bar graph: gaps

histogram: no gaps unless a class has a count of 0

$0 < \leq 4$ 10

$4 < \leq 8$ 0

$8 < \leq 12$ 5

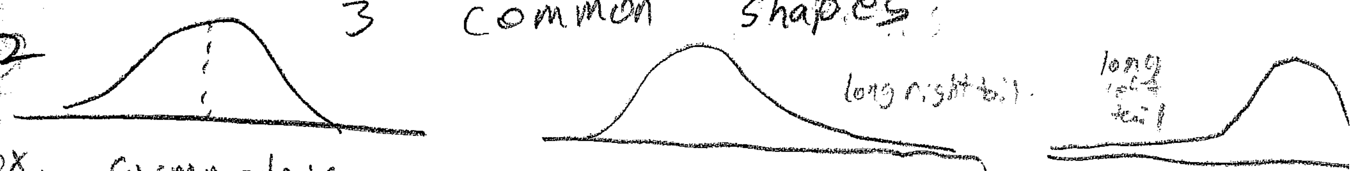


8) p10 Since the bars have equal width, 282 – 2
the area of each bar is proportional to the percentage of individuals in each class.
1st paragraph p10 is not important

9) p11 To interpret the histogram,
Look for an overall pattern and deviations from the pattern. shape, center, spr[illegible]

10) p11 <u>Outliers</u> are observations that fall outside the overall pattern.

11) p12 3 common shapes:



approx. symmetric — approx. mirror image about midpoint — Eg 1.3 p13

right skewed — eg income data — Eg 1.4 p13

left skew[ed] — eg scores from easy exam

12) Know Stemplots p15 are for small data sets (p17)

Final exam scores

Stems	Leaves
3	0 4 5 6 7 8
4	2 6 6
5	0 1 2 4 7 8 9 9
6	0 1 4 4 4 5 5 5 6 7 8 8 8 9
7	0 1 1 2 2 2 3 3 4 5 6 6 8 8 8 8 8 9
8	1 1 1 2 2 2 2 3 4 4 4 5 6 7 7 7 8 8 8
9	1 2 2

↑ Stems — leaves

Stem ten[s], leaf one[s]

six[ties], thir[ties], [illegible], fort[ies], eighte[ies], ninete[ies]

a) leaf = final digit

divide data into groups = stems that contain all but the last digit
eg for exam data 3, 4, ..., 9 stand for scores in 30's, 40's, ...

b) Place stems in order in a vertical column (~~text~~ suggest smallest on top largest on bottom) Draw a vertical line to separate stems from leaves

c) write each leaf to the right of its stem in increasing order

d) write stem and leaf units on plot (not in text)

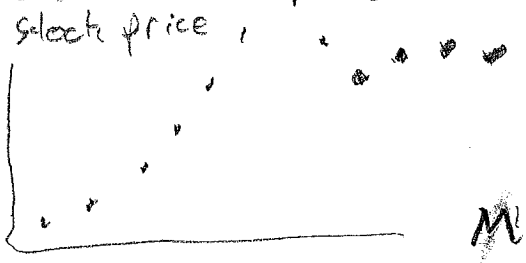
13) p ~~17~~ variations on the stem plot

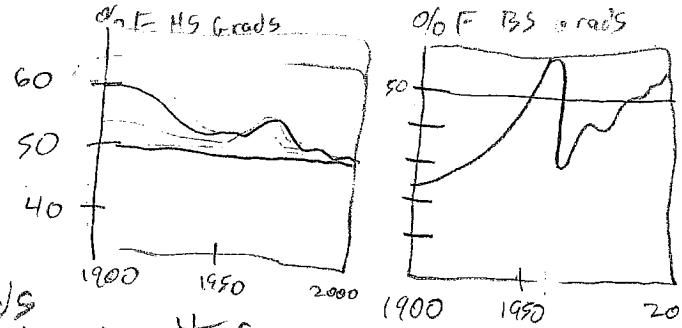
a) If there are too many stems, round the data, then make a stem plot

b) split the stems eg 7 0-4
7 5-9

one stem for 70, 71, 72, 73, 74
another stem for 75, 76, 77, 78, 79

14) p17 a <u>time plot</u> plots each observation, vs the time it was measured





19 brands

ex) fat contents of peanut butter

22, 30, 40, 56, 69, 30, 41, 50, 65,
39, 40, 56, 62, 36, 45, 50, 44, 53, 56

2	2
3	0 0 6 9
4	0 1 0 4 5
5	6 0 0 3 6 6
6	8 5 2

scratch

0282 - 3

2	2
3	0 0 6 9
4	0 0 1 4 5
5	0 0 (3) 6 6 6
6	2 5 8

stem ten
leaf one

5(10) + 3(1) = 53

split stems 0-4

hardly ever use this

2	2
2	
3	0 0
3	6 9
4	0 0 1 (4)
4	5
5	0 0 3
5	6 6 6
6	2
6	5 8

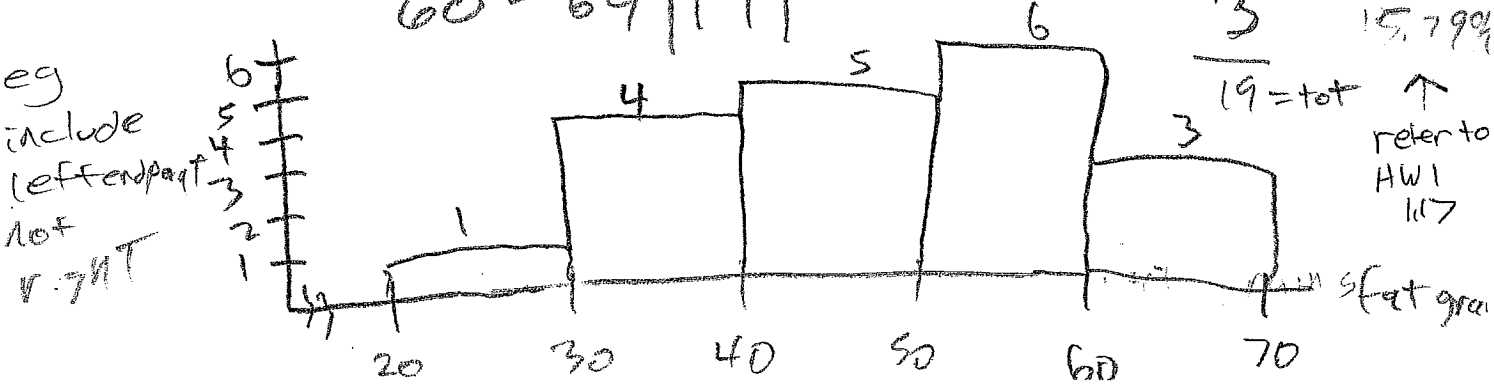
4(10) + 4(1) = 44

stem tens
leaf ones

% = $\frac{\text{count}}{\text{tot}}$ 100%

Histogram

	tally	Count	%
20 - 29	1	1	5.26
30 - 39	1 1 1 1	4	21.05
40 - 49	1 1 1 1	5	26.32
50 - 59	1 1 1 1 1	6	31.58
60 - 69	1 1 1	3	15.79
		19 = tot	



refer to HW1 1.17

15) p 37 mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum x_i$ memorize
Know

ex) 1, 9, 3, 7 $\quad \bar{x} = \frac{1+9+3+7}{4} = \frac{20}{4} = 5$

Note: could say find $\bar{y}$, $\bar{w}$, $\bar{x}$. The letter should not matter.

16) know p 30 median M

a) sort the data from smallest to largest

b) count $\frac{n+1}{2}$ obs's from the bottom of the list. (n is the number of obs's)

- If n is odd $M = \frac{n+1}{2}$ th observation
  even M = average of the 2 middle obs

Note: at least half of the obs's $\geq M$ and at least half are $\leq M$.

ex) 1, 9, 3, 7 $\xrightarrow{\text{sort}}$ 1, 3, 7, 9 $\quad$ note $\frac{n+1}{2} = \frac{5}{2} = 2$

n = size of list = 4

$M = \frac{3+7}{2} = 5$

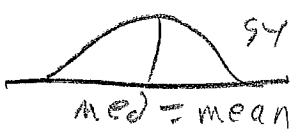
ex) 1, 9, 3 $\xrightarrow{\text{sort}}$ 1, 3, 9 $\quad \frac{n+1}{2} = \frac{3+1}{2} = 2$

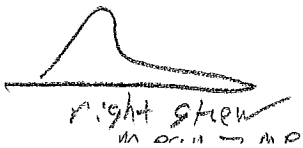
↑ 2nd obs

$M = 3$

17) $\bar{x}$ and M are measures of <u>center</u>

18) p 49 If a distribution is (exactly) symmetric the median = mean





02824

19) the range = largest obs − smallest obs

20) p33 Order the list and find the median. Then the 1st quartile Q1 is the median of the obs's to the left of the median M while the 3rd quartile Q3 is the median of the obs's to the right of the median M. The 2nd quartile $Q_2 = M$

ex) n even 50, 59, 72, 81 ← ordered

$Q1 = 54.5$ $Q3 = 76.5$

ex) n odd 50, 59, 72, 81, 97 ← ordered

$Q1 = 54.5$ $Q3 = \frac{81+97}{2} = 89$

$Q2 = M = 72$

ex 4 3

4, 7, 7, 7, 8, 9, 9

↑ Q1 ↑ Q2 ↑ Q3

21) know p35 Box plot

a) order the data; find the five number summary min Q1 M Q3 max

b) Make a box from Q1 to Q3. Draw a line for M.

c) Draw a line from the edge of the box (Q1) to min.

Draw a line from the other edge of the box (Q3) to the max, label axes

ex exam scores (stem plot) min = 30

Q1 = med of 1st 35 obs = 18th obs = 60

M = 72 (36th obs) Q3 = med of last 35 obs = 54th obs = 82 max = 92

	min	Q1	M	Q3	max
scores	30	60	72	82	92

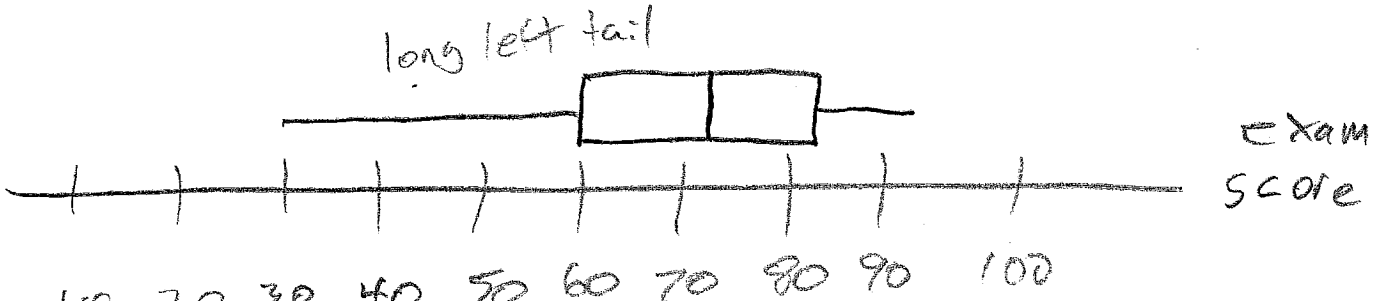
1 2 ... 18 ... 35 | 36 37 ... 54 ... 71

$\frac{35+1}{2} = 18$ (18th obs)

37 = 36+1

36+18 = 54 (54th obs)

71 = 36+35



22) p38 The standard deviation (know) memorize

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

The SD measures spread.

Note: The variance $= s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$

ex) use table for HW, exam, Quiz

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	−2	4
2	−1	1
3	0	0
4	1	1
5	2	4

$\Sigma x = 15$

$\bar{x} = \frac{15}{5} = 3$

$0 = \Sigma(x-\bar{x})$ good check, p39

$\Sigma(x-\bar{x})^2 = 10$

$= (x_1-\bar{x})^2 + \cdots + (x_5-\bar{x})^2$

$$s = \sqrt{\frac{10}{5-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.581$$

common mistakes 1) Forget square root (2.5) ($\sqrt{10} = 3.16$)

2) forget to divide by $n-1$

23) **common Q E problem**

Find $\bar{x}$, median and s for

5, 4, 1, 2, 3 ($\bar{x} = 3$, $M = 3$, $s = 1.58$)

p39 $x_i - \bar{x}$ is called a deviation $\Sigma(x_i - \bar{x}) = 0$

24) properties p39 s measures spread about $\bar{x}$

$s \geq 0$ and $s = 0$ only when all obs's have the same value. s is larger when there is more spread

s has the same units as the original obs's

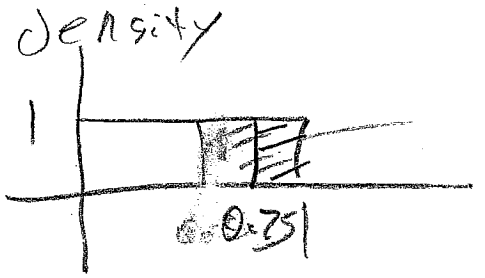
25) A dot graph is the best description of a distribution; always plot first

Ø 1.3

26) p48 A density curve is never negative and the area under the curve = 1.0

Areas under curve represent proportions of observations

ex)



shaded area = [1 − (0.75)] 1 = .25

25% of x values are between 0.75 and 1

Density curves are often useful approximations for histograms

27) p49-50 The median of a density curve is the point such that half of the area is to the left and half the area to the right.

The mean μ of a density curve is the balance point. See figure 1.16.

→ The mean is drown towards the longer tail in skewed histograms.

incomes

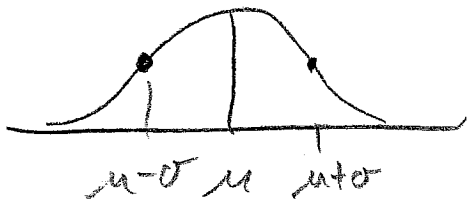
28) p54 σ = SD of density curve

$\bar{x}$ estimates μ

s estimates σ

29) p51 the normal curve is a density curve that can approximate many data sets. Given μ and σ, a normal curve can be drawn. A normal curve is symmetric about μ, so μ is the mean and the median.

see fig 1.19 p52



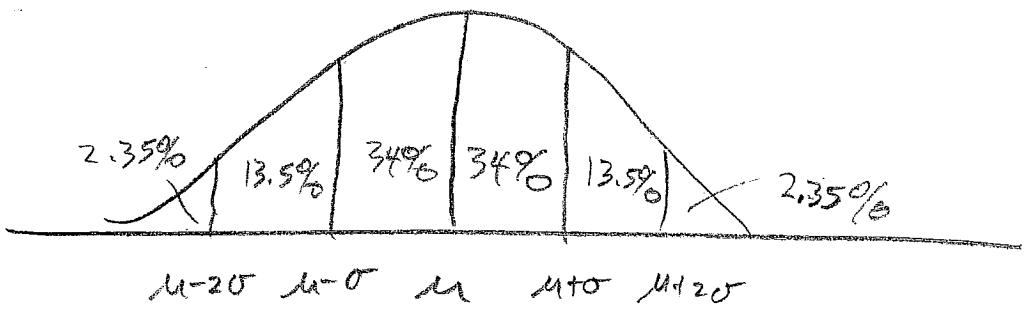
30)* p53 68-95-99.7 rule

For normal data

68% of obs's are within σ of μ

95% 2σ

99.7% 3σ



ex length of rattlesnake is approx normal with $\mu = 8$ in $\sigma = 1$ in Then $\approx$ 95% have lengths between 6 in and 10 in. $\approx$ 2.5% are longer than 10 in

31) Notation x from $N(\mu, \sigma)$ means x is an obs from a normal dist with mean μ and SD σ.

32) P52 Last paragraph on p52 is important

3 reasons for normal

many datasets $\simeq$ normal

many are not

33) know P55 If x is from a dist with mean μ and SD σ, then the standardized value of x is the

z score of x $= z = \frac{x-\mu}{\sigma}$.

and tells how many SD's x is from μ. (memorize)

34) P56 If X is $N(\mu, \sigma)$, then

$z = \frac{X-\mu}{\sigma}$ is $N(0,1)$, the standard normal. Hence one table can be used for every normal curve.

35) P57-61 The z table in the front of the book can be used to find 3 types of

Probabilities (areas).

Let Z be $N(0,1)$. The z-table tables $P(Z \le z^*) = P(Z < z^*)$.

I) a) $P(Z < 2.8)$

z^*	.00
2.8	.9974

b) $P(Z \le -1.57)$

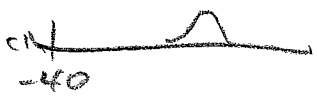
z^*	.07
-1.5	.0582

c) $P(Z \le -4)$

$< P(Z \le -3.49)$

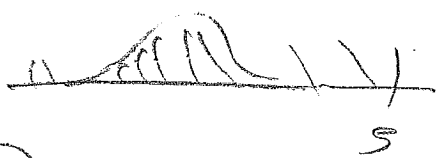
so $P(Z \le -4) \approx 0$.

d) $P(Z \le -40) \approx 0$



e) $P(Z \le 5) > P(Z < 3.49)$

so $P(Z \le 5) \approx 1.0$



f) $P(Z \le 0) = 0.5$

Key words: less than

II $P(z > z^*) = P(z \geq z^*)$

$P(z > z^*) = 1 - P(z \leq z^*)$

$P(z \leq z^*)$ — area from table

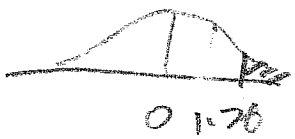
a) $P(z \geq -1) = 1 - P(z < 1) =$

$= 1 -$

$= 1 - .1587 = .8413$

b) $P(z \geq 1.763) = P(z \geq 1.76)$ ← closest

$= 1 - P(z < 1.76) = 1 - .9608 = .0392$



c) $P(z \geq 0) = 0.5$

d) $P(z > 1.645)$

1.64 and 1.65 are equally close to -1.645

$= 1 - P(z < -1.645)$

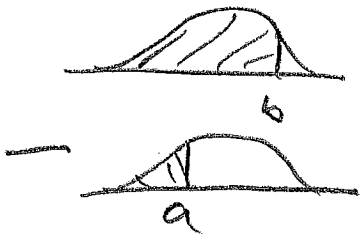
$= 1 - \left(\frac{.0505 + .0495}{2}\right)$

$= 1 - .05 = 0.95$

key words! greater than

III $P(a \le z \le b) = P(a < z \le b) = P(a \le z < b)$
$= P(a < z < b)$

$= P(z \le b) - P(z \le a)$  =



a) $P(-1 \le z \le 1) = .8413 - .1587 = .6826$

b) $P(2.25 \le z \le 4.09) = P(z \le 4.09) -$

$P(z < 2.26) \quad = 1 - .9881 = 0.0119$

Key words "between."

30) p 271 Finding Probabilities when x is N(μ,σ)

Forwards Know for E Final:

step1 state problem with a picture, 2) Standardize, 3) draw picture
4) use z table

i) $P(x < b) = P(x \le b) = P\left(z \le \frac{b-\mu}{\sigma}\right)$ ← table

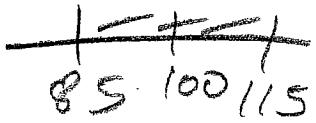
ii) $P(x > a) = P(x \ge a) = P\left(z \ge \frac{a-\mu}{\sigma}\right) = 1 - P(z \le$ tab

iii) $P(a \le x \le b) = P(a < x < b) = P(a \le x < b) = P(a < x \le$

$= P\left(\frac{a-\mu}{\sigma} < z \le \frac{b-\mu}{\sigma}\right) = P\left(z \le \frac{b-\mu}{\sigma}\right) - P\left(z \le \frac{a-\mu}{\sigma}\right)$ table

ex) IQ test scores are $\sim N(\mu=100, \sigma=15)$

a) Find chance $85 < X < 115$

step1 

$= P(85 < x < 115)$

step 2 standaradize

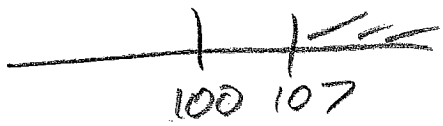
$\frac{85-100}{15} = -1$ $\frac{115-100}{15} = 1$

step3 

$= P(-1 \le z \le 1)$

step4 $.8413 - .1587 = .6826$

b) Find chance randomly selected IQ score is greater than 107

step1 

$= P(x \ge 107)$

step2 standardize x=107 $z = \frac{107-100}{15} = 0.47$

step3 

$= P(z > 0.47)$

step4 $= 1 -$ 

$= 1 - .6808 = .3192$

common mistake standardize, then write down table value (eg .6808 instead of .3192)

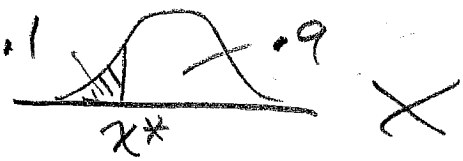
37) p 61 Backwards Normal calculations: Finding an x value given a proportion

Z 82 09 know for Exam Fact

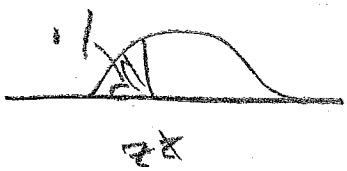
step 1 state problem with a picture step 2 use table

step 3 unstandardize

ex) For the IQ scores find the score needed to be in the highest 90%

step 1 

Find x^* so that $P(x \le x^*) = .1$

step 2 

Find the largest prob $\le .1$ and the smallest prob $\ge .1$

Then z^* corresponds to the closer

z^*	.08	.09
−1.2	(.1003)	.0985

closer

so $z^* = -1.28$

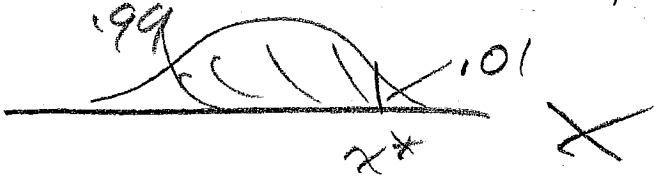
Step 3 unstandardize $z^* = \frac{x^* - \mu}{\sigma}$

so $x^* = \mu + z^*\sigma$ p72

$x^* = 100 + 15(-1.28) = 100 - 19.8 = 80.2$

Find the IQ score x^* such that 99% of the scores are smaller, (q.5)

Step 1



Step 2

z^*	.02	.03
2.3	.9898	.9901 closer

$z^* = 2.33$

Step 3 Unstandardize

$x^* = \mu + z^* \sigma$

$= 100 + 15(2.33) / 15 = 134.95$

check Standardize x^*

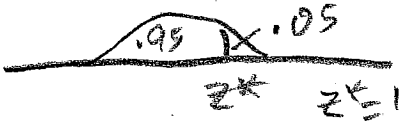
$$\frac{134.95 - \mu}{\sigma} = \frac{134.95 - 100}{15} = 2.33$$

$= .9901$

---

ex) Suppose women heights $\sim N(\mu = 66\text{in}, \sigma = 3\text{in})$

Find height so that 5% are taller



$z^* = \frac{1.64 + 1.65}{2} = 1.645$

.04	.05
.9495	.9505

$x^* = \mu + z^* \sigma = 66 + 3(1.645) = 70.93$

Ch 2

0282 10

1) p [illegible] a response variable

= dependent variable Y. measures an outcome of a study. An explanatory variable = independent variable X attempts to explain the observed outcomes.

p [illegible] sometimes have 2 variables, but they are not response explanatory

ex predict Y for given values of X

eg Y = college Fresh GPA X = SAT score

Y is the response X the explanatory variable

2) p [illegible] 135 association ≠ causation

2 variables X and Y can be related, but this does not mean that changes in X result in changes of Y

ex X = shoe size Y = reading score of kids 4-8

bigger x's go with bigger y's

increase in age results in increase of shoe size and reading score

3) p82 A scatter plot is a plot of X vs Y

eldest son ht = Y

dad ht = X

Y = brain wt

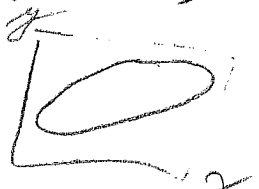
horizontal axis

vert axis

age = X

4) p84 Look for an overall pattern especially linear (football shaped) plots and for outliers

5) p84 2 variables are positively associated when above ave values of one variable tend to accompany above ave values of the other and below ave values also tend to occur together

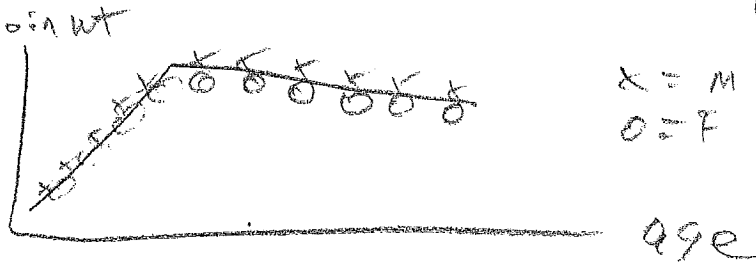


2 variables are negatively associated when above ave values of one variable tend to accompany below ave values of the other and vice versa.

p85 strong relationships make accurate predictions possible.

6) p84 an outlier is an observation that falls outside the overall pattern. (see p18, too)

7) p88 use colors or symbols to show categorical variables in a scatterplot of 2 quantitative variables



8) p99 The strength of a linear relationship of 2 quantitative variables x and y is measured by the correlation r.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{1}{n-1}\sum z_x z_y$$

where $\bar{x}$ and s_x are the mean and SD of x
while $\bar{y}$ and s_y are the mean and SD of y,

ex — common exam problem delete 5 numbers

x husband	y wife	$\frac{x-\bar{x}}{s_x}=z_x$	$\frac{y-\bar{y}}{s_y}=z_y$	product $z_x z_y$
72	66	1.1858	0	0
68	64	−0.3953	−0.9535	0.3769
70	66	0.3953	0	0
68	65	−0.3953	−0.4767	0.18843
71	70	0.7905	1.9069	1.50740
65	65	−1.5810	−0.4767	0.75366
		0.0000 always true except for rounding	0.0000	2.82639
$\Sigma x = 414$	$\Sigma y = 396$			$\Sigma z_x z_y$

$\bar{x} = \frac{414}{6} = 69$ $\quad \bar{y} = \frac{396}{6} = 66$ $\quad 72-69=3$

$$s_x = \sqrt{\frac{1}{n-1}\Sigma(x_i-\bar{x})^2} = \sqrt{\frac{1}{5}\left[3^2+(-1)^2+1^2+(-1)^2+2^2+(-4)^2\right]}$$

$$= \sqrt{\frac{32}{5}} = \sqrt{6.4} = 2.530$$

$$s_y = \sqrt{\frac{1}{n-1}\Sigma(y_i-\bar{y})^2} = \sqrt{\frac{1}{5}\left[0^2+(-2)^2+0^2+(-1)^2+4^2+(-1)^2\right]}$$

$$= \sqrt{\frac{22}{5}} = \sqrt{4.4} = 2.0976$$

3rd column, 1st row: $\frac{72-69}{2.530} = 1.1858$

4th column, 1st row: $\frac{65-66}{2.0976} = -0.4767$

Each row last column $= (-1.5810)(-0.4767) = 0.75366$

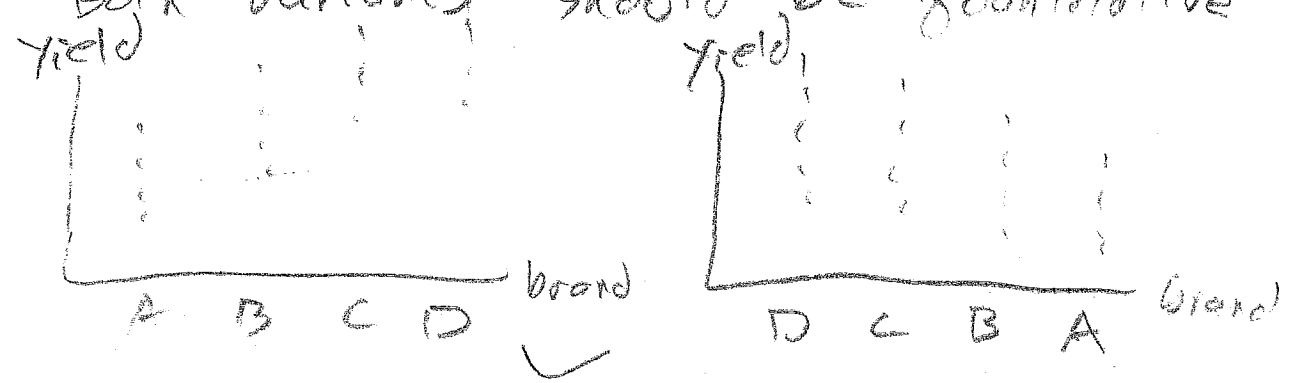
Since $\sum z_x z_y = 2.82639$,

$$r = \frac{2.82639}{5} = 0.565$$

exam Q: get table with 4 entries missing a z_x, z_y and products

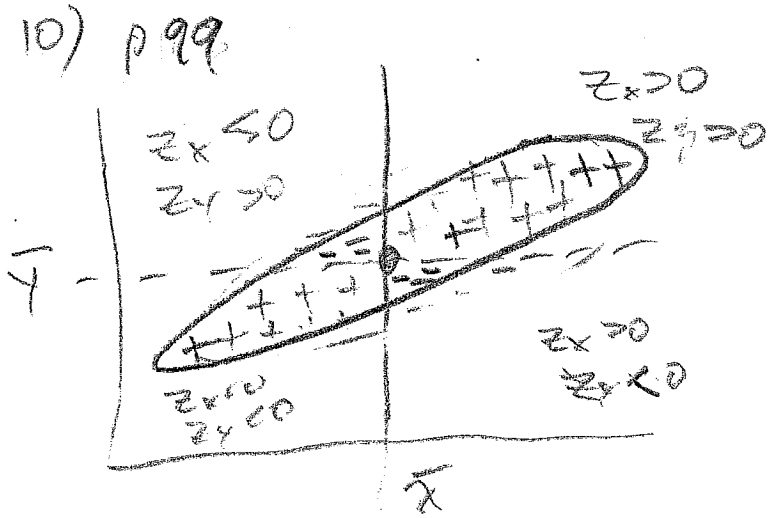
9) p98 Both variables should be quantitative

ex.

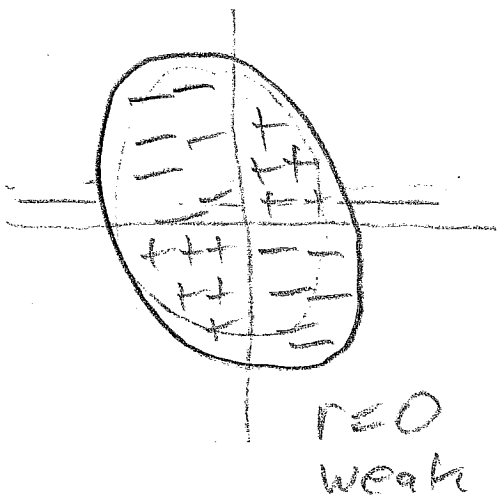


positive and negative association doesn't make sense

10) p99



r>0 strong



r=0 weak

11) p99 Properties of r when Scatterplot is (football shaped) linear

corr x,y = corr y,x, unit free

$-1 \le r \le 1$

positive assoc if $r > 0$

neg $<$

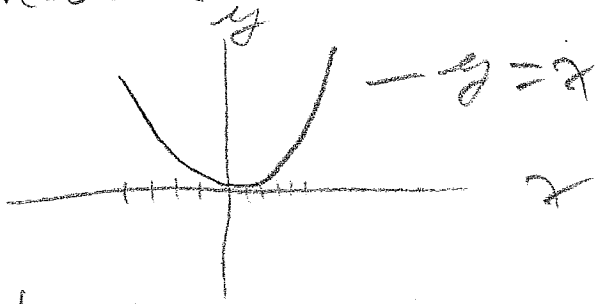
r far from 0 means strong association

ie knowing x helps a lot in predicting y

$r = \pm 1$ if the points in the scatterplot fall exactly on a line

∴ adding or subtracting the same # to all values of x or y does not change r

12) If the scatterplot is nonlinear, r should not be used



perfect nonlinear relationships but $r = 0$ is possible

ϕ 2.3

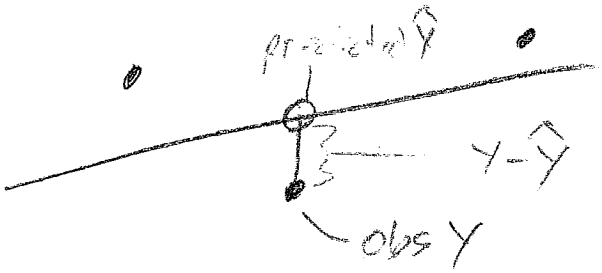
13) p106 The regression line describes how the response variable y changes as the explanatory variable x changes.

p109

14) A line has the form $y = a + bx$ where b = slope, a = intercept. The least squares line picks a and b so that $\Sigma(y - \hat{y})^2$ is minimized where

$a = \bar{y} - b\bar{x}$, $b = r\left(\frac{s_y}{s_x}\right)$ (s_y = SD y, s_x = SD x, r = corr x,y) $\hat{y} = a + bx$ (predicted value)

$y - \hat{y}$ = vertical error



15) LS line $\hat{y} = a + bx$, $a = \bar{y} - b\bar{x}$

$b = r\frac{s_y}{s_x}$. know for Q Exam Final

Given $\bar{y}, \bar{x}, r, s_y, s_x$, you should be able to get the LS slope b and intercept a.

16) p123 the slope b is the amount $\hat{y}$ changes when x changes by one unit

17) To plot the LS line take one of the smallest values of x and find $\hat{y}$, " largest "

These 2 points determine the line.

18) p131 Extrapolation is using regression to predict y for x outside of the range of the observed values of the explanatory variable, such predictions are often very bad.

ex) Predict wt from ht

men 18-24 x = ht y = wt

$\bar{x} = 70$ in $s_x = 3.$ in $r = 0.47$

$\bar{y} = 162$ lb $s_y = 30$ lb

min ht = 57 max ht = 79

$b = 0.47\frac{30}{3.} = 4.7\frac{lb}{in}$

$a = \bar{y} - b\bar{x} = 162 - 4.7(20) = -167.16$ 0282 13

$\hat{y} = a + bx = -167 + 4.7x$ is the LS line

Predict y if i) $x = 20$: $\hat{y} = -167 + 4.7(20) = 162$ lbs

ii) $x = 73$: $\hat{y} = -167 + 4.7(73) = 176.1$

iii) $x = 60$: $\hat{y} = -167 + 4.7(60) = 115$

iv) $x = 1$: $\hat{y} = -167 + 4.7(1) = -162.3$

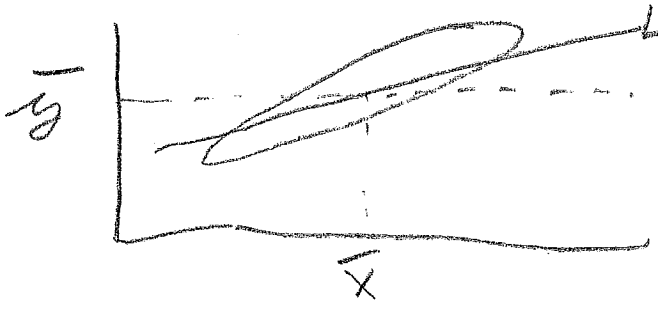
19) p113 Units of x and y matter

slope $= b = r\frac{s_y}{s_x}$ means a change in 1SD in x results in a change of r SD in y

$(-1 \le r \le 1)$

The LS line always passes through $(\bar{x}, \bar{y})$.

20) p113 The square of the correlation, r^2 is the fraction of the variation of y (about $\bar{y}$) that is explained by the LS line



The line $y = \bar{y}$ corresponds to $b = 0$, $a = \bar{y}$

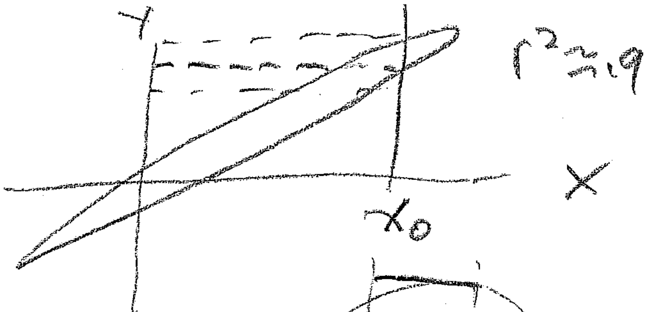
$\frac{1}{n-1}\sum (y-\bar{y})^2 = s_y^2 \le \frac{1}{n-1}\sum (y-\hat{y})^2$

LS minimizes this

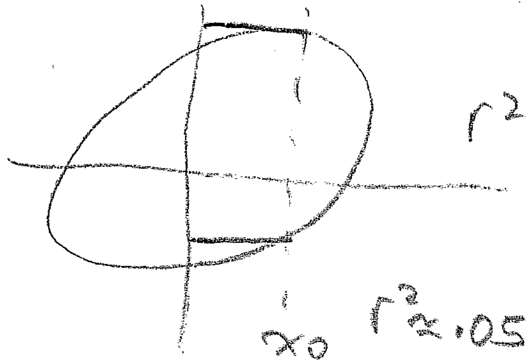
If there is a linear relationship r^2 small means that to predict a new value of y, $\bar{x}$ is almost as good as LS, i.e. knowing x

does not help much in predicting y.

r^2 large means the ls line is much better than $\bar{Y}$ for predicting a new y value: knowing x helps a lot in predicting y.



If r^2 is close to 1, for fixed x_0, spread in y is small (compared to s_y).



r^2 close to 0, for fixed x_0, spread in y is large (close to s_y)

ex $Y = \frac{9}{5}X + 32$ $Y = F^0$ $X = C^0$ $\quad x = \frac{5}{9}(Y - 32)$

suppose $Y = 10 \quad 50 \quad 90$ then $\bar{Y} = \frac{150}{3} = 50$ $\quad \bar{Y} = 10, 50, 90$

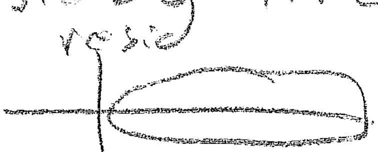
$X = -\frac{110}{9} \approx -12.22 \quad 10 \quad \frac{290}{9} \approx 32.22$ $\quad$ at $X = -\frac{110}{9}, 10, \frac{290}{9}$

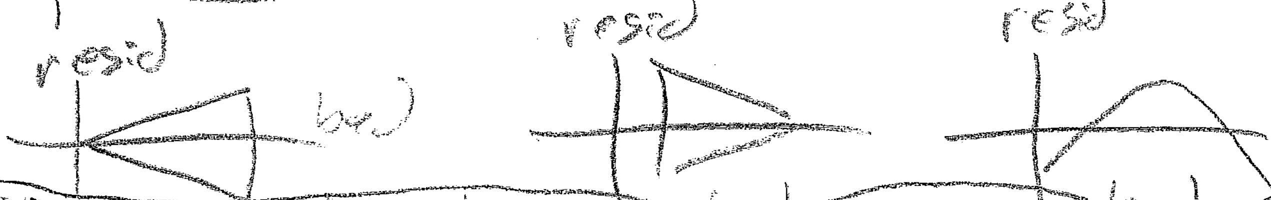
21) p116 residual $= Y - \hat{Y}$

p128 a <u>residual plot</u> is a scatterplot of the residuals vs x or of the residuals vs $\hat{Y}$

want to see residuals scattered about resid = 0 line with no pattern

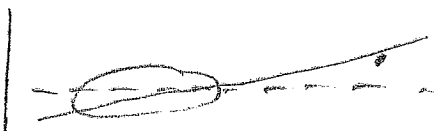
Fact $\Sigma(Y - \hat{Y}) = 0$





22) p123 ... influential if deletion ...

22) p124 an observation is <u>influential</u> if deleting it and then refitting the LS line results in a marked change in the fit.



Minitab
know for Q Exam Final
gas used = 1.09 + .188 ddays

23) Minitab output p2.4

Minitab output p110

	coef	
pred constant	1.0892 = a	predict y if x = 20
x → D days	0.1888 = b	$\hat{y} = 1.0892 + .1888(20)$
		$= 4.9652$

24) p130 r and LS can be very misleading if the relationship between x and y is nonlinear, eg if outliers are present.

25) p132 A <u>lurking variable</u> p xxvi is a variable that has an important effect on the relationship among the variables in the study but is not included among the variables studied.

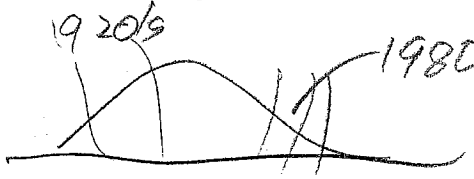
skip ex shoe size = x y = reading score of kids 2–8

age of child is a lurking variable

skip ex NY Times Feb 24 1998

Some IQ tests have been given over and over again eg every year by the military. Each year they are standardized to have mean 100 and SD 15. A researcher got year by year results and the raw scores

The IQ scores have been going up each year.



Conclusion

generation X is far more intelligent than their grandparents generation

X ~~Year~~ IQ test taken in military

Y = test score

possible lurking variables:

1) literacy rate increased
2) military more selective, no draft
3) Standardized tests are converted to IQ's so schools teach ~~a~~lot more IQ test problems than they did.

ex) Raw SAT, ACT scores decreased from 1950 to 1990

conclusion educators ability decreased

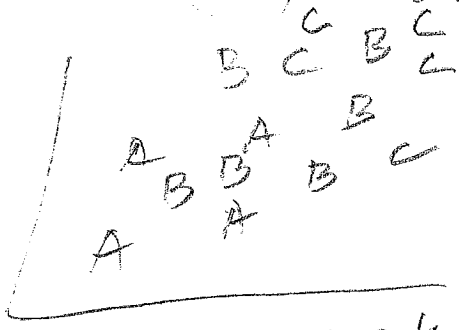
lurking 1) 50's top 5% students took exam now top 2/3 take it

2) Immigration demographics have changed. Perhaps fewer % of immigrants have English as a native language,

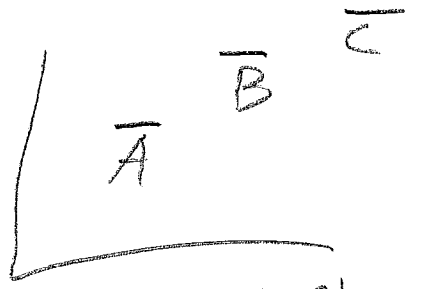
26) p134 Often averages are computed
men 35-54 1990 census
eg Y = ave income 9 regions of USA
X = ave education by 9 regions

corr = 0.7

but among individuals
corr (x,y) $\simeq$ 0.4



corr weak



corr strong

Correlations based on averages are usually higher than correlations based on individuals (the correlation that is usually of interest).

skip 2-7) p135 Association does not imply causation,
eg shoe size reading score kids 2-8

ch 3

1) p168 A population is the entire group of individuals we want information about. A sample is the subset of the population actually studied.

2) p168 The design of the sample is the method used to choose the sample from the pop.

3) p169 p xxviii A voluntary response sample consists of people who choose themselves by responding to a general appeal.
^ ... sample chooses the individuals easiest to reac

ex) Magazines often print surveys 15.
eg on sexual behavior, and ask readers to respond.
problem! People who respond tend to differ from the pop as a whole

ex Ann Landers pxxviii 10000 letters 70% said kids not worth it. This biases the sample.
A statistically designed poll said 91% would have kids again.

ex) 1936: The Literary Digest, had correctly picked the winner in the past 5 presidential elections. It sent questionaires to 10 million people for Landon vs FDR presidental election. They got the names from phone books, club lists, etc
The digest poll said Landon would win overwhelmingly 57% to 43%
In 1936 on 25% of homes had phones.

Gallop took a sample of 3000 people from the same lists and predicted the Digest results before the Digest published (off by 1%), with a sample of 50000

Gallop predicted	FDR 56%	Langdon 44%.
Actual result	62%	38%
Digest	43%	57%.

After the election, Digest's cover said "Boy are we red in the face."

~~The Digest poll was not a voluntary response sample since they chose who got~~ 10-82 16
the ~~questionaires~~. The sample had a much larger proportion of wealthy than the pop as a whole. Before 1936 rich and poor had similar voting behavior. In 1936 the poor esp 9 million unemployed were for FDR the rich (eg 11 million phone owners) were for Langdon.

Of those who got questionaires, only 20% responded. Respondants tend to differ from non respondants

4) p174 A probability sample gives each member of the population a known nonnegative probability (chance) of being selected.

5) p180 The result from a probability sample, eg the proportion of Republicans that will vote for Bush, will follow some density curve, often the normal distribution. eg probability sample of 1000 Republicans .627 say they will vote for Bush $\approx N(\mu = .627, \sigma = .015)$

So 95% of samples of size 1000 will give a proportion between $\mu \pm 2\sigma$ or .597 to .657

With a probability sample, you can give a guarantee on how accurate the result is. 16

p180 This guarantee is called the "margin of error."

6) p170 A probability method uses impersonal chance to select the sample

7) Polling companies such as Gallup now use probability samples (the design).

Many surveys have terrible designs

p169 a) voluntary response sample

b) samples of convenience: ask the people easiest to reach (the interviewer chooses the sample)

ex1) ask SIU student passing through the student center if they are Rep or Dem

ex2) At a conference on Leukemia give questionaires to Doctors regarding improvements in treatment. (Interviewer chose all people attending although many won't respond)

8) p169 The design is biased if it systematically favors certain outcomes.

In ex1) Too many women got asked

In ex2) doctors who attend conferences differ from doctors who do not.

Digest poll got to many rich.

ex) Send person out to 10 different neighborhoods, have person interview people at 3 homes per neighborhood. Wealthier are over represented, so Poorer will have an equal

are under represented)

9) p171 The simplest probability sample is a simple random sample (SRS).

For a SRS of size n from a pop of size N, the prob that any member of the pop gets in the sample is $\frac{n}{N}$. That is, each person has an equal chance of being selected.

Idea! put everyones name in a hat, shuffle the names, then draw n out.

ex) pop of size 10,000 sample size n = 1000
Put # 1 – 10,000 with each name. Use random numbers to select the 1000,
chance any person is selected = $\frac{1000}{10000} = \frac{1}{10}$.
Every sample of size 1000 out of 10000 (an enormous #) is equally likely.

ex) pop size = 7 a b c d e f g
1 2 3 4 5 6 7
table B line 130 sample of size 2
6 9 0 5 ef

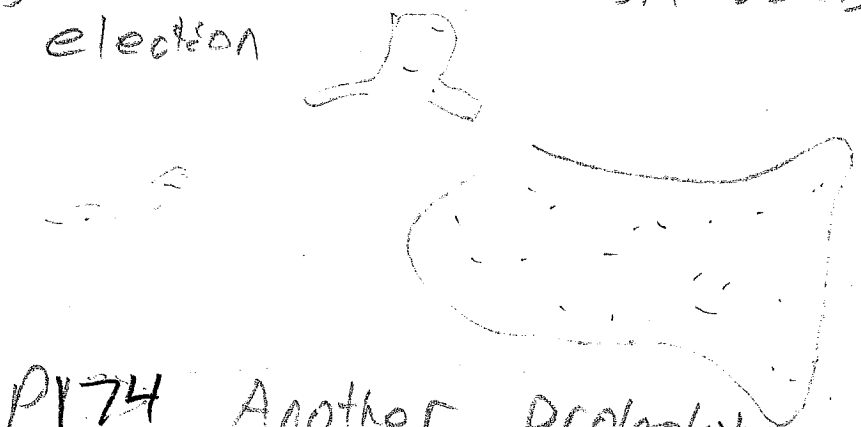
ex) pop = 31 want sample of size 5

table B: Start at line 130 $a_{01}, a_{02}, \ldots, a_{31}$

69 05 16 48 17 87 ~~47~~ 40 95 ~~17~~ 84 53
40 64 89 87 20 19

persons $a_5, a_{16}, a_{17}, a_{40}, a_{20}$ (in order selected)

10) Problem with SRS. They are too expen 17.3
if the pop is big and spread out.
eg SRS of 1000 USA voters for 2000 pres election

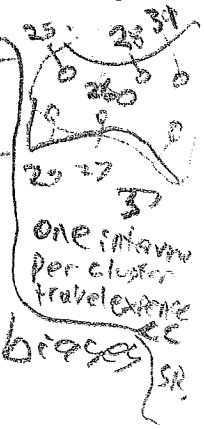


11) p174 Another probability sample is a stratified random sample. First divide pop into groups of similar individuals called strata,
eg by age race gender income and education.
Take a SRS from each strata to get the full sample.

12) p173 A probability method called a multistage sample is used if interviewers are sent out. Idea Regions → SRS of pop centres → SRS of wards → SRS of households → ask every voter in the selected households
eg current population survey → SRS of precincts → SRS of households

The design yields clusters of nearby households. One interviewer can visit each cluster, so travel expense is much lower than a SRS of households

13) Random digit dialing. Regions are determined by area code), then use random phone #'s,



14) Probability samples eliminates many biases but there are still problems

15) p178 Undercoverage Some groups 282 78

eg homeless, prison inmates, college students tend to be underrepresented (eg get sample proportions and compare to census proportions)

16) p178 Nonresponse Individual chosen for sample can't be contacted or won't cooperate (non respondents differ from respondents)

Want survey response rate to be high eg > 80% (Digest only had 20% response rate).

17) p178 response bias

respondents may say what they think the interviewer wants to hear

eg attitude of interviewer!

18) wording of question

Do you use tide?

What detergent do you use?

Show me the brand of detergent you use.

19) Census controversy

Census wants to use prob methods to [illegible]

20) p181 For the accuracy of a sample depends on the sample size, (provided that the sample size is a small percentage of the pop size), really if pops differs $\frac{\sigma_1}{\sqrt{n_1}} \leq \frac{\sigma_2}{\sqrt{n_2}}$ means sample 1 is more accurate and if both samples are from the same pop

ex) Find proportion of women

SRS of 2500 Carbondale residents

SRS 2500 Chicago residents

The results will have the same accuracy.

ex) Ask everyone in class whether R or not.

Get 100% accuracy if you ask everyone.

ex) .01% Alaskans are polled and .01% CA are polled which sample is more accurate or are the accuracies the same?

Soln CA sample is much larger so more accurate

1) p214 A <u>parameter</u> is an unknown number that describes a <u>population</u> eg μ or σ.

A <u>statistic</u> is a number computed from the <u>sample</u> (without any knowledge of the parameters) that estimates a parameter. eg $\bar{x}$, s

2) p215 The sample proportion $\hat{p}$ estimates the pop proportion p. eg 1000 voters in sample, 511 Dem

$\hat{p} = \frac{511}{1000} = 0.511 = 51.1\%$

3) p215 <u>Sampling variability</u> different samples produce different values of the statistic.

The chance behavior of a statistic has a regular and predictable pattern if the process (experiment) that produces the statistic is repeated many times (long run behavior),

ex coin toss expt 1 H, 2 T, 3 H, 4 H — $\frac{3}{4}$ H's but in long run expect $\frac{1}{2}$ H's

		# dem	$\hat{p}$
ex Sample	1	511	.511
	2	522	.522
	3	507	.507
	4	517	.517

ch10 ↓

Use many samples and make a histogram of the $\hat{p}$'s Hist $\approx N(p, \sigma_{\hat{p}})$

4) p217 A phenomenon is <u>random</u> if individual outcomes are uncertain but in the long run follows a distribution like the normal. eg flip coin, toss dice

ex] $Y = X^2$ if $x = 2$ $y = 4$ not random

grade on exam, spin of roulette wheel

5) p217 The <u>probability</u> of an outcome of a random phenomenon is the proportion of times the outcome would occur in a long series of independent experiments.

Figure 1: 4 DD Plots

eg flip fair coin prob Heads $= \frac{1}{2}$

6) p218 independent means the outcome of one trial (expt) must not influence the outcome of any other trial

eg) new manufacturing employee learns so # of mistakes decreases. If we count her <u>mistakes per day</u> the trials are not independent. trial

7) p218 Simulation ← ch10

ex) generate many samples with a computer for each sample generated compute a statistic eg $\hat{p}$'s Make a histogram of $\hat{p}_1, \ldots, \hat{p}_{1000}$. Find mean and SD of $\hat{p}_1, \ldots, \hat{p}_{1000}$ etc.

<u>ex</u> Simulate fair coin Prob(H) = 0.5 ← ch10

assign 0-4 to H 5-9 T toss die $P(2) = \frac{1}{6}$

1: 1 9 2 2 3 9 5 0 $\hat{p} = \frac{5}{8} = 0.625$

H T H H H T T H

2: 5 4 5 8 0 8 1 5 $\hat{p} = \frac{3}{8} = 0.375$ etc.

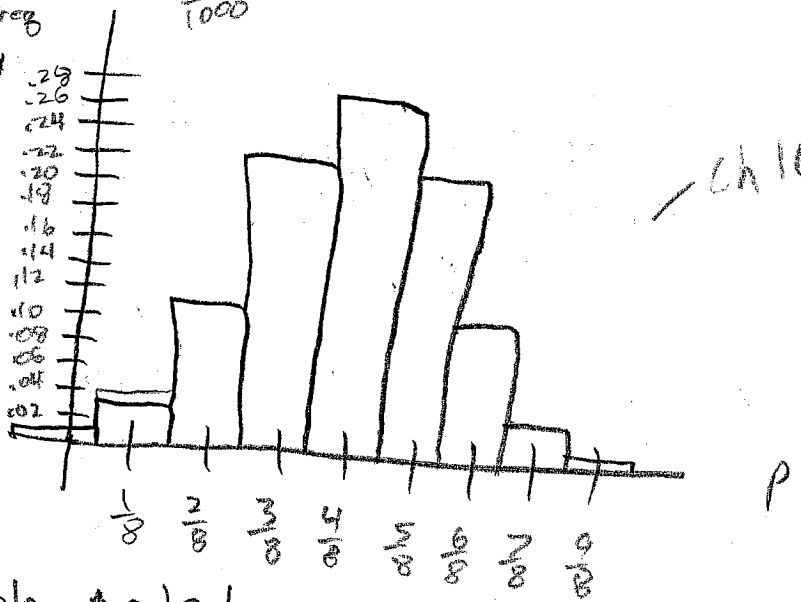
T H T T H T H T H

Generate 1000 samples

# heads	count	freq
0	4	.004
1	31	.031
2	109	.109
3	219	.219
4	273	.273
5	219	.219
6	110	.110
7	31	.031
8	4	.004
	1000	

$\frac{4}{1000} = .004$ etc

0282 20



Ch 10

§4.2

§4.2 8) p220 prob model Sample space S

= set of all possible outcomes.

An event A is a subset of the sample space S.

A probability model consists of S and a way of assigning probabilities to events.

ex toss coin once $S = \{H, T\}$ $P(H) = P(T) = \frac{1}{2}$

ex genders 2 child families $S = \{bb, bg, gb, gg\}$ (1st, 2nd)

event 2 boys $\{bb\}$

prob .5 b .5 g so each of the 4 outcomes has prob $= \frac{1}{4}$,

prob [at least one boy] $= P\{bb, bg, gb\} = \frac{3}{4}$

9) p223 **rules of prob**

1) The prob $P(A)$ of any event A satisfies $0 \le P(A) \le 1$

2) $P(S) = 1$ where S is the sample space

3) complement rule $P(A \text{ does not occur}) = 1 - P(A) = P(\text{not } A)$

A and B are disjoint $P(A \text{ or } B) = P(A) + P(B)$ 20

A or B mean A or B or both occur

If $A_1, A_2, \ldots, A_k$ are disjoint, $P(A_1 \text{ or } A_2 \cdots \text{or} A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$

addition rule for disjoint events

ex) grades

A	B	C	D	F
.2	.3	.2	.15	.15

are disjoint $P(S) = .2 + .3 + .2 + .15 + .15 = 1.0$

$P(\text{C or higher}) = .2 + .3 + .2 = .7$

$P(\text{D or lower}) = .15 + .15 = 1 - .7 = .3$

ex) Die toss 2 fair die

list all possibilities using order

eg 1st toss, 2nd toss

S given on p 221

1st \ 2nd	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2						
3						
4						
5						
6						(6,6)

etc

each outcome has prob $\frac{1}{36}$

(0) P225 Finite $S = \{s_1, s_2, \ldots, s_n\}$ $P_1 \; P_2 \qquad P_n$ $P_k = P(s_k)$

i) $0 \le P_k \le 1$ and ii) $\sum_{i=1}^{n} P_i = 1$

each outcome in S has a prob between 0 and 1 and the sum of all the probs equals 1

iii) $P(A) \triangleq \sum_{\text{outcomes in } A} P(\text{outcomes in } A)$

ex) die A = sum of 2 die $= 7$ $P(A) = P(1,6), (2,5), (3,4), \ldots, (6,1) = \frac{6}{36} = \frac{1}{6}$

ex) toss fair coin 3 times

list all possibilities using order of toss →

1st	2nd	3rd
H	H	H
H	H	T
H	T	H
H	T	T
T	H	H
T	H	T
T	T	H
T	T	T

8 outcomes in S each with prob $\frac{1}{8}$

ex)

	A	B	C	D	E
Prob	.1	.2	.3	.1	

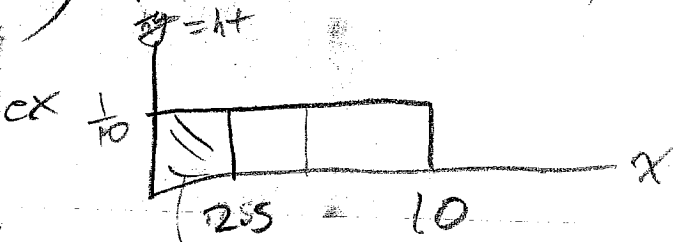
$P(F) = 1 - (.1 + .2 + .3 + .1) = .3$

Since the prob's sum to 1

11) If outcomes are equally likely, $P(A) = \frac{\#\text{ outcomes in } A}{\#\text{ outcomes in } S}$

12) Prob from density curve

ex



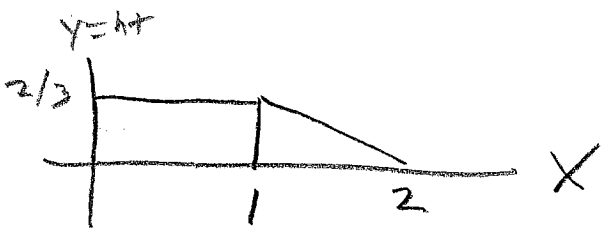
rect area = base(ht) = 1

$10\left(\frac{1}{10}\right) = 1$

$P(X < 2.5) = 2.5\,\frac{1}{10} = .25$

triangle area = $\frac{1}{2}$ base(ht)

ex



$1(ht) + \frac{1}{2}\,1(ht) = 1$

so $\frac{3}{2}ht = 1$ $ht = 2/3$

$\text{Prob}(X > 1) = \frac{1}{2}\,1\left(\frac{2}{3}\right) = \frac{1}{3}$

$\text{Prob}(X < 1) = 1 - \frac{1}{3} = 1\left(\frac{2}{3}\right) = \frac{2}{3}$

ex) normal curve

13) p231 a random variable X is a variable whose value [illegible] numerical outcome of a random phenomenon

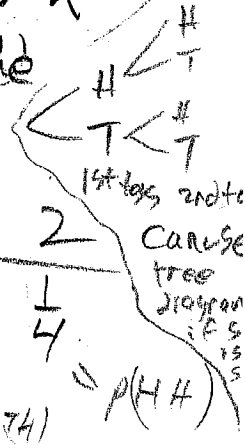
eg toss coin 2 times X = # heads $S = \{0, 1, 2\}$

(BK) p231 The probability distribution of RV X tells the possible values of X along with the probs of those values.

HH HT TH TT

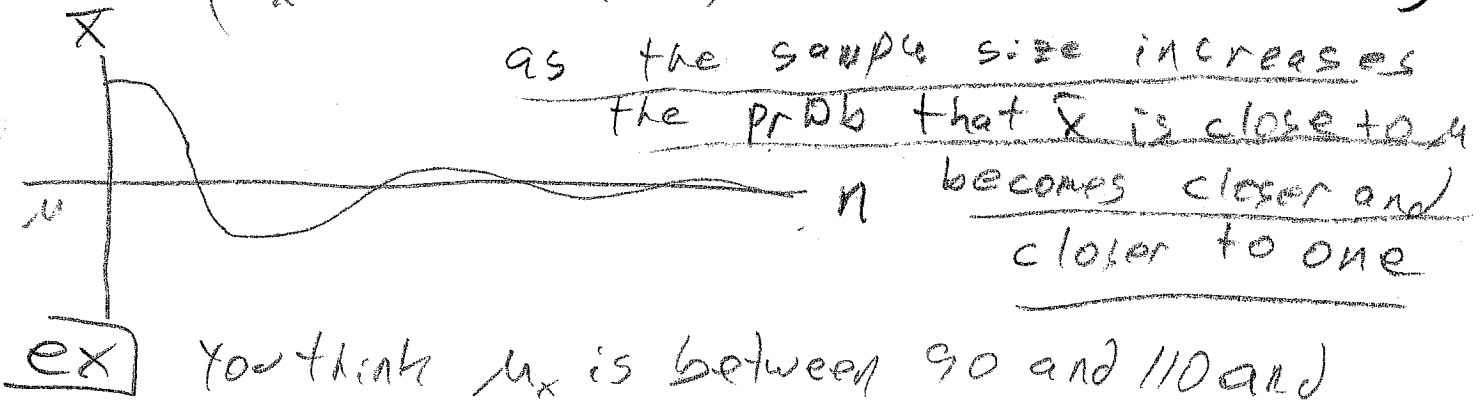
$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$

X	0	1	2
Prob	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	‖ P(TT)	‖ P(HT) + P(TH)	= P(HH)



§4.3 15) p237 **Law of large numbers**

Suppose you want to estimate μ_X, the mean of X with $\bar{X} = \frac{\sum x_i}{n}$, where the x_i are drawn at random. (Given any tolerance value, say $\delta > 0$, (for large enough n) $P(\mu_X - \delta < \bar{X} < \mu_X + \delta)$ is arbitrarily close to 1.)



as the sample size increases the prob that $\bar{X}$ is close to μ becomes closer and closer to one

ex You think μ_X is between 90 and 110 and want to estimate μ_X to within $\pm .001$ with 99% probability. Then there is some integer N so that for any randomly drawn sample of size $n \geq N$

$P(\mu_X - .001 \leq \bar{X} \leq \mu_X + .001) \geq 0.99$

Implications i) Casino gambling: If you make many bets that are as small as allowed, you will lose a lot of money.

Bet as much money as you can in as few bets as possible to maximize your chances of winning a lot of money (but you will still probably lose)

ii) Stock market stocks tend to increase

Buy small amounts of lots of shares and your portfolio should behave like the stock market average. If you put all your money in one stock, your gains or losses could be huge.

16) p241 The sampling distribution of a statistic is the distribution of the [illegible] of the values taken by the statistic in all possible samples of the same size) = pop of statistic

	variable	population	mean	SD
ex	x	pop of x	μ	σ
	$\bar{x}$	sampling dist of $\bar{x}$ = pop of $\bar{x}$	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$

ex pop net worths of 3 people Bill Gates 90 Billion
MJ 0.8 B You 0.0 B Find the sampling distribution of $\bar{x}$ for samples of size 2

		$\bar{x}$
BG	MJ	$90.8/2 = 45.4$
BG	Y	$90/2 = 45$
MJ	Y	$0.8/2 = 0.4$

$\bar{x}$	45.4	45	0.4
prob	1/3	1/3	1/3

17) p239 x comes from an underlying population with μ_x σ_x. A statistic, eg $\bar{x}$, comes from a pop = sampling distribution of the statistic which also has a center and spread eg $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$.

18) p243 typically the center of the sampling distribution is close to what the statistic estimates (eg μ_x for $\bar{x}$) and the spread of the sampling distribution decreases as the sample size n increases. See fig 4.10 p242.

19) p242 A statistic is an unbiased estimator for a parameter if the mean of the sampling distribution of the statistic is equal to the parameter.

$\bar{x}$ is unbiased for μ_x $\hat{p}$ for p

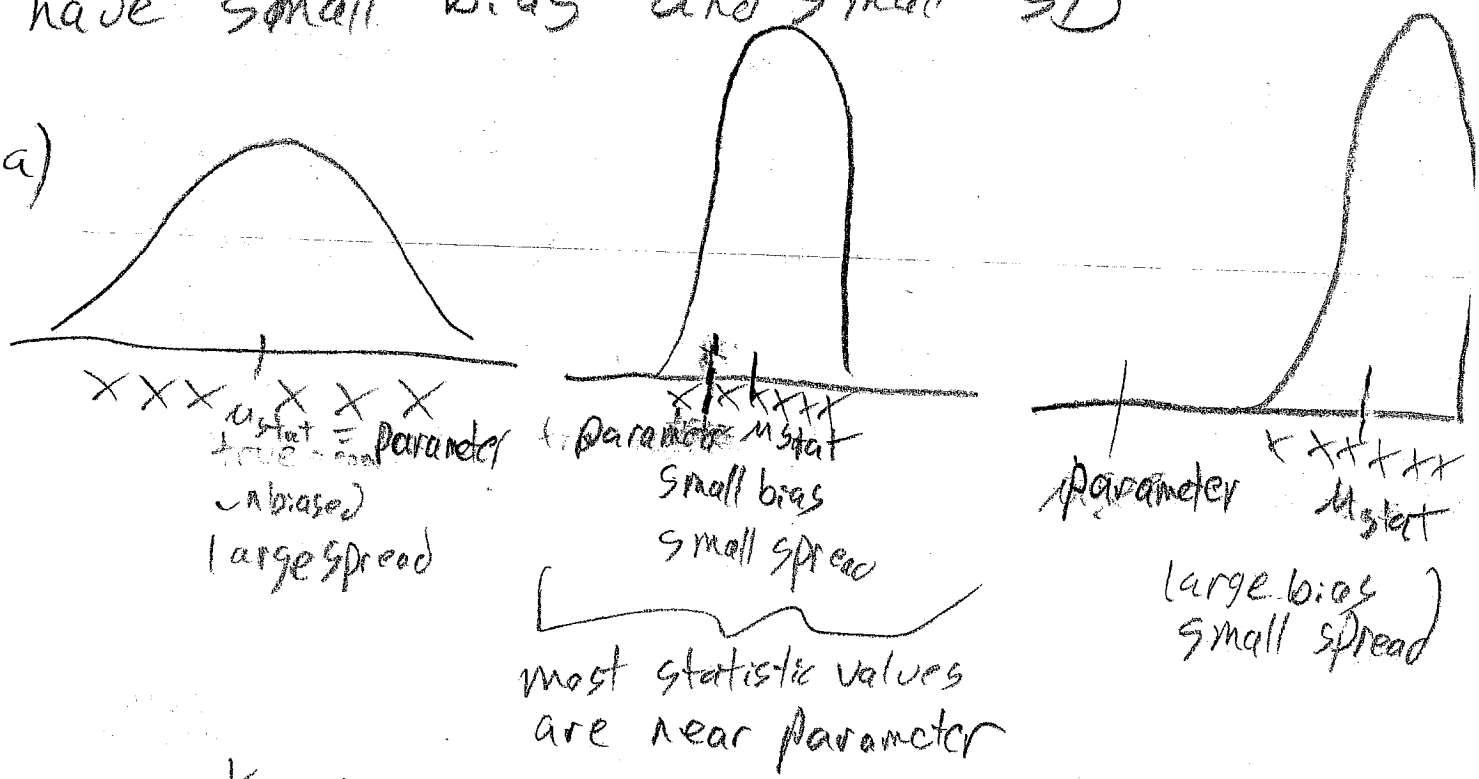
s^2 is unbiased for σ^2

c is not unbiased for

20) p239 The variability of a statistic is described by the SD of the sampling dist of the stat. For many statistics, larger samples have a smaller SD than smaller samples.

eg $\sigma_{\bar{x}_{1000}} < \sigma_{\bar{x}_{100}}$, Often the sampling dist SD does not depend on the pop size provided that the pop size is a lot larger than the sample size.

21) Want the sampling distribution to have small bias and small SD



22) p242, 244 Know The sampling distribution of $\bar{x}$ obtained from a SRS of size n from the population of x having mean μ_x and SD σ_x has

i) mean $\mu_{\bar{x}} = \mu$

ii) SD $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ if $n \leq$ pop size / 10

iii) If x is $N(\mu_x, \sigma_x)$ then $\bar{x}$ is $N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$.

iv) Central limit theorem (CLT) If n is large and x is from any distribution with finite σ_x and mean μ_x then

$\bar{x} \simeq N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$.

Note: i) implies that $\bar{x}$ is unbiased for μ_x

ii) implies that $\bar{x}$ is less variable than x if $n \geq 2$

iii) implies $\bar{x}$ is normal if x is normal

iv) implies that $\bar{x}$ is approx normal if n is large

23) The results for $\bar{x}$ also hold if the n observations come from n experiments done under identical conditions. This is like a SRS from the pop of all possible measurements (expt results),

Luke Tierney
2000-02-14

24) Producing a sample with randomization ! the impartial use of chance, "washes out the effects of lurking variables, minimizes biases, and can make the sampling dist of the statistic easy to approx, eg normal

So $\bar{X}$ is efficient for $\tau(\theta) = \frac{\theta}{\theta-1} - \frac{1}{\log\theta}$

7.37

Done in class.

2.9 How large should n be to use the CLT?

~~F 28 28~~ use 0282 24

i) $n \geq 1$ for X normal know

ii) $n \geq 5$ for X close to normal

iii) If X has highly skewed underlying POP, dont use the normal approx if $n \leq 30$, know

iv) If $n \geq 100$ CLT usually holds on this class.

Usually I will say "X is normal" (so $\bar{X}$ is too) or "assume n is large enough so that the CLT holds" exception iii)

ex See fig 4.11 p 245 common exam problem

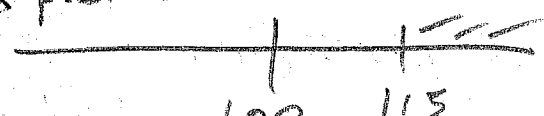
$n = 100$ X_i = ith IQ score

$\approx$ normal with $\mu = 100$ $\sigma = 15$

i) Find $P(\bar{X} > 115)$

Step 0 ↓ Key step ↓

Step 1 $\bar{X}$ picture



$\bar{X}$, $\mu_{\bar{X}} = 100$ $\sigma_{\bar{X}} = \frac{15}{\sqrt{100}} = 1.5$

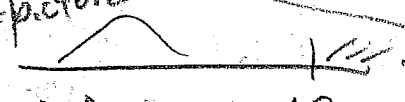
triplet, get from POP, not from sampl

Step 2 Standardize $Z = \frac{115 - 100}{1.5} = 10$

$= \frac{\text{x bar value} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$

Zscore is always from a triplet; $\frac{\text{value} - \text{mean}}{\text{SD value}}$

Step 3 Z picture

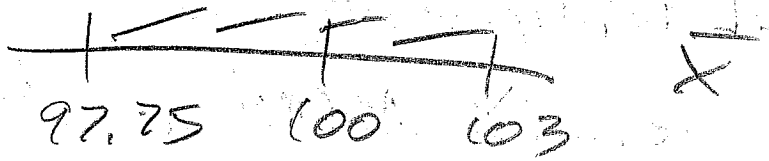


Step 4 table A

$P(Z \geq 10) = 1 - P(Z \leq 10) = 1 - 1 = 0$

11) Find $P(97.75 < \bar{x} < 103)$ 24.5

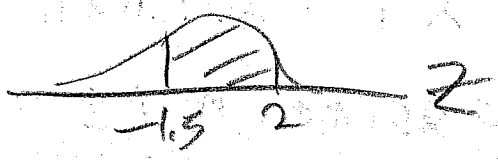
Step 1



Step 2 $z = \frac{97.75 - 100}{1.5} = -1.5$

$\frac{103 - 100}{1.5} = 2$

Step 3



Step 4 $= .9772 - .0668 = .9104$

C $P(-1.5 < z < 2) = P(97.75 < \bar{x} < 103)$

← likely exam, quiz question

ex a) Suppose x comes from a highly skewed distribution $n=10$, $\mu=100$, $\sigma=15$. Find $P(\bar{x} \geq 115)$ if possible.

and x comes from a normal dist with

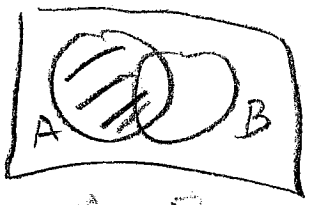
b) If $n=10$, find $P(\bar{x} \geq 115)$ if possible.

soln a) CLT does not apply, can't do it

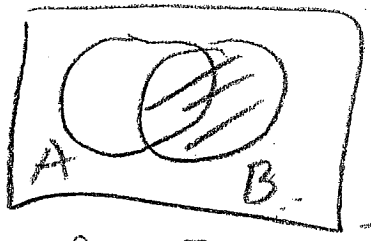
b) C 100 115 $\bar{x}$, $\mu_{\bar{x}}=100$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.74$ $z = \frac{115-100}{4.74} = 3.16$

$= 1 - .9992 = .0008$

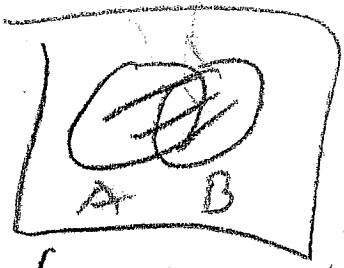
Ch 5 1) p259 Venn diagrams 0282 25



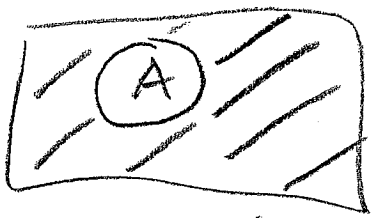
for A



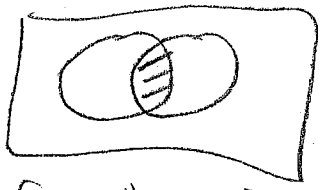
for B



for A or B



for not A



for A and B

2) p260 2 events A and B are independent if knowing that one occurs does not change the prob that the other occurs

n events A1, ..., An are independent if knowing that any subset of 1 to n-1 events occurs does not change the probs of the others.

3). If events are not independent, then they are dependent.

ex If $0 < P(A), P(B) < 1$ and A and B are disjoint, then A and B are dependent since knowing A occurred means B did not occur (an extreme form of dependence)

ex Identical experiments done under identical conditions are independent eg coin toss!

common sense often works

height & IQ

hair color and height

Your final grade coin toss

25.5

4) Two events A and B are independent if $P(A \text{ and } B) = P(A)\,P(B)$ and vice versa

5) Multiplication rule p260

If A and B are independent $P(A \text{ and } B) =$ P(both events A and B occur) $= P(A)\,P(B)$.

If events $A_1, A_2, \ldots, A_n$ are independent, then

$P(A_1 \text{ and } A_2 \text{ and } \ldots A_n) =$ P(all n events occurred)

$= P(A_1)\,P(A_2) \ldots P(A_n)$ (*) but if $n > 2$,

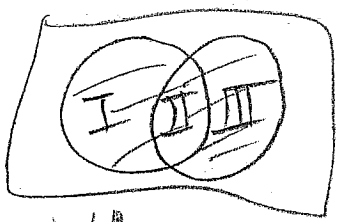
(*) does not imply that $A_1, \ldots, A_n$ are independent.

6) p264 General addition rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

note $P(A) + P(B) = P(A \text{ or } B) + P(A \text{ and } B)$

double counted



I A and not B
II A and B III B and not A

so $P(A \text{ or } B) = P(\text{I}) + P(\text{II}) + P(\text{III}) = P(A) + P(B) - P(\text{II})$

7) Summary i) $0 \le P(A) \le 1$

ii) $P(S) = 1$

iii) complement rule $P(\text{not } A) = 1 - P(A)$

iv) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ addition rule

v) If A and B are disjoint, then
$P(A \text{ and } B) = 0$ while $P(A \text{ or } B) = P(A) + P(B)$ ← add rule for disjoint events

vi) If $A_1, A_2, \ldots, A_n$ are disjoint, then
$P(A_1 \text{ or } A_2 \text{ or } \ldots A_n) = P(A_1) + \ldots + P(A_n)$

vii) If A and B are independent, $P(A \text{ and } B) = P(A)P(B)$
and $P(A \text{ or } B) = P(A) + P(B) - P(A)P(B)$.

viii) If $A_1, A_2, \ldots, A_n$ are independent then
$P(A_1 \text{ and } A_2 \ldots \text{and } A_m) = P(A_1)P(A_2)\ldots P(A_m)$

ix) $P(A) = \sum P(\text{outcomes in } A)$ if S is finite

x) List all possibilities using order of trials
eg roll 2 die, find sum toss coin 3 times, find # heads

6) p262-263 combining prob rules
Perform n identical experiments.
Interest is in a single outcome of the expt, say D
Find i) P(D occurred in none of the n experiments)
ii) P(D occurred in at least one of the n experiments)
iii) P(all n outcomes were D)
iv) P(not all outcomes were D)

i) $P(\text{none}) = \underbrace{\frac{1-P(D)}{1}}_{\text{comp rule}} \frac{1-P(D)}{2} \ldots \frac{1-P(D)}{n} = \underbrace{[1-P(D)]^n}_{\text{mult rule}}$

ii) P(at least one) = P(not none) = 1 − P(none)

Read ex 5.3 and Sec 5.4 p 262-263
Cover all!

UNIVERSITY OF MINNESOTA
Twin Cities School of Statistics

Luke Tierney
1999-12-13

ex 70 rolls: $P\{\text{none are 5's}\} = \frac{1-\frac{1}{6}}{1^{st}} \cdot \frac{1-\frac{1}{6}}{2nd} \cdots \frac{1-\frac{1}{6}}{70th} = (\frac{5}{6})^{70}$

i) $P\{\text{not all 3 rolls are 5's}\} = 1-(\frac{1}{6})^3$

ii) $P\{\text{all 3 rolls are 5's}\} = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} = (\frac{1}{6})^3$
1st 2nd 3rd

3 none

at least one

$P(\text{at least one}) = 1 - P(\text{none})$

iii) $P(\text{at least one of the 3 rolls is a 5}) = 1-(\frac{5}{6})^3$

i) $P(\text{none of the rolls are 5's}) = \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6} = [1-\frac{1}{6}]^3$

ex Roll die 3 times D = die was 5 in the n^{th} toss

iv) $P(\text{not all}) = 1 - P(\text{all})$

iii) $P(\text{all}) = \underset{\text{def of rule}}{=} \frac{P(D)}{1^{st}} \cdot \frac{P(D)}{2nd} \cdots \frac{P(D)}{n^{th}} = [P(D)]^n$ (2)

common exam prob's

ex $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ 0 2>

4 given day 3 -- find the remaining

eg A and B are ind $P(A) = .2$ $P(B) = .6$
Find $P(A \text{ or } B)$.

Soln $P(A \text{ and } B) = .2(.6) = .12$

$P(A \text{ or } B) = .2 + .6 - .12 = 0.68$

ex $P(A) = .6$ $P(B) = .5$ are A and B disjoint or is more information needed?

Soln if A and B are disjoint

$P(A \text{ or } B) = P(A) + P(B) = .6 + .5 = 1.1$ impossible

So A and B are not disjoint

ex 100 students 40 men 60 women
5 men with beards randomly select a person

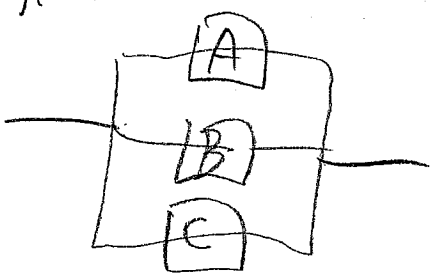
$P(A) = P(\text{man with beard is selected}) = \frac{5}{100} = .05$

$P(B) = P(\text{man selected}) = \frac{40}{100} = .4$

$P(A \text{ and } B) = P(A) = \frac{5}{100} \neq P(A)P(B)$

So A and B are dependent

skip ex i)



device fails if all components fail

exii) thief tries to break in house there are 3 triggers for alarm A B C

P(device fails) = P(trigger goes off) = .01 .02 .03

Prob	exi)	exii
.01 .02 .03 = ≈0.000	all components fail = device fails	thief set off all 3 triggers (very stupid thief)
.99 .98 .97 = ≈ .94	none of the components fail	thief does not set off any of the triggers alarm does not go off
1 − .01 .02 .03 = ≈ 1.000	at least one component does not fail = device runs	at least one trigger does not go off
1 − .99 .98 .97 ≈ .06	at least one component fails	thief fails to disable at least one trigger = alarm goes off

↑ skip

§5.2 9) Let X be a count (of "successes")

eg take a SRS of size $n=100$ and 57 are F then $X = 57$ if F was counted

the proportion $\hat{p} = \frac{X}{n}$ so $X = n\hat{p}$

10) p269 The distribution of X can be derived if i) there are n observations

ii) these are independent

iii) each observation is a 1, if it is a "success" (what you want to count) 0 if not a success. Hence there are 2 categories success and not a success

iv) prob (success) = prob (1) = p

prob (0) = 1 − p

11) p270 If i) – iv) hold, the distribution of X is called a binomial distribution with parameters n and p. X is bin(n, p).

X	0	1	2	...	n	
prob	p_0	p_1	p_2	...	p_n	$p_i =$ P(X = i)

12) p271 The count of X from a SRS is approx bin(n, p) if the sample size n is small compared to the pop size.

So $h(t) = \frac{1}{\beta}F(t)$.

UNIVERSITY
OF MINNESOTA
Twin Cities School of Statistics

Luke Tierney

1999-12-13

15) p 273 $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ ~~F282 26~~ 028229

for $k = 0, 1, \ldots, n$.

Idea: If $X = k$ then there were k success and $n-k$ failures.

$$\underbrace{S}_{1}\ \underbrace{S}_{2}\ \cdots\ \underbrace{S}_{k}\ \underbrace{F}_{k+1}\ \underbrace{F}_{k+2}\ \cdots\ \underbrace{F}_{k+n-k}$$

By independence $P(S \cdots S\ F \cdots F) = \overbrace{[P(S)]^k}^{p}\ (P(F))^{n-k}$ ($\leftarrow 1-p$)

(A and B ind $\Rightarrow P(A \text{ and } B) = P(A)P(B)$)

But there are $\binom{n}{k}$ ways to choose the k S's among the n outcomes

so $P(X=k) = \binom{n}{k} (p)^k (1-p)^{n-k}$. $x^0 = 1, x \neq 0$ Fact $p^0 = 1$, $(1-p)^0 = 1$

16) p274 If X is bin n p then

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

ex) a machine produces light bulbs. 80% meet specifications. A sample of 6 bulbs is taken from the machine's production. If 3 or more fail to meet specifications, the

Machine will be recalibrated. (29.

Find the probability that the machine will be recalibrated.

Soln want $P(\text{at least 3 are defective})$

Let X count the number of defectives

X is bin $n=6$ $p=.2$

Want $P(X \geq 3) = P(X=3) + P(X=4) + P(X=5) + P(X=6)$

↑ add by p 244, the 4 events have no outcomes in common

$$= \left[\binom{6}{3}(.2)^3(.8)^3 + \binom{6}{4}(.2)^4(.8)^2 + \binom{6}{5}(.2)^5(.8)^1 + \binom{6}{6}(.2)^6(.8)\right]$$

$$= 20(.008)(.512) + 15(.0016)(.64) + 6(.00032)(.8) + 1(.000064)$$

$$= 0.08192 + .01536 + .001536 + .000064$$

$$= 0.09888 \approx 10\%$$

$$\mu_x = 6(.2) = 1.2$$

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{6(.2).8} = \sqrt{.96} = .9798$$

~~#45 a) Describe the sampling distribution of $\bar{X}$ obtained from a SRS of size n from the population of X having mean~~

Next: Week 9 **Up:** STATISTICS 8101Homework Assignments **Previous:** Week 7

Luke Tierney

1999-12-13

17) Relationship with p 261–263

× counts number of successes

$P(X=0) = P(\text{none}) = \binom{n}{0} p^0 (1-p)^n = (1-p)^n$ (none were successful)

$\frac{1-p}{1} \quad \cdots \quad \frac{1-p}{n}$

$P(X=n) = P(\text{all}) = \binom{n}{n} p^n (1-p)^0 = p^n = \frac{p}{1st} \cdots \frac{p}{nth}$ successes

$P(\text{at least one success}) = P(X \geq 1) = 1 - P(X=0) = 1 - (1-p)^n$

$P(\text{not all}) = P(X \leq n-1) = 1 - P(X=n) = 1 - p^n$

18) at most and at least

$P(\text{at least } j \text{ successes}) = P_j + P_{j+1} + \cdots + P_n = P(X \geq j)$

$P(\text{at most } j \text{ successes}) = 1 - P_0 - P_1 - \cdots - P_{j-1}$

$P(\text{at most } j \text{ successes}) = P(0) + P(1) + \cdots + P(j) = P(X \leq j)$

$= 1 - P_n - P_{n-1} - \cdots - P_{j+1}$ where $P_k = P(X=k)$

ex 20 multiple choice questions 50 students

Each each. Student randomly guesses

Let X = # questions answered correctly/

$n = 20$ always adds to n $p = \frac{1}{5} = .2$

i) $P(X=16) = \binom{20}{16} (.2)^{16} (.8)^4$

$= \frac{20!}{16!\,4!}(.2)^{16}(.8)^4 = 4845(.2)^{16}(.8)^4 = 1.3 \times 10^{-3}$

P (Student gets at most 2 correct) = $P(X \le 2)$

$= P_0 + P_1 + P_2 = \binom{20}{0}(.2)^0(.8)^{20} + \binom{20}{1}(.2)^1(.8)^{19} + \binom{20}{2}(.2)^2(.8)^{18}$

8 of 10 points

$= (.8)^{20} + 20(.2)(.8)^{19} + 190(.2)^2(.8)^{18}$

$= .011529 + .057646 + .136908$

$= .2062$

iii) P (student gets at least 3 correct)

$= P(X \ge 3) = 1 - P_0 - P_1 - P_2$

$= 1 - .2062 = .7938$

$= 1 - \binom{20}{0}(.2)^0(.8)^{20} - \binom{20}{1}(.2)^1(.8)^{19} - \binom{20}{2}(.2)^2(.8)^{18}$

18) p 278 Suppose X is $b(n,p)$ (and $n \le$ pop size / 10 usually true).

If $np \ge 10$ and $n(1-p) \ge 10$

then $X \approx N\left(\mu_x = np,\ \sigma_x = \sqrt{np(1-p)}\right)$

20) If $np < 10$ or $n(1-p) < 10$ don't use the normal approx. know

ex] 20 multiple choice questions OM2§2 3/

p = prob student gets correct answer $= \frac{4}{5}$

knew .75 guessed on .25 if took test many times over and over again

i) If $n=20$ can normal approx be used?

$np = 20\frac{4}{5} = 16 \geq 10$ $\quad n(1-p) = 20\frac{1}{5} = 4 \leq 10$

NO

ii) If $n=100$ find the prob student gets at least 85 questions correct.

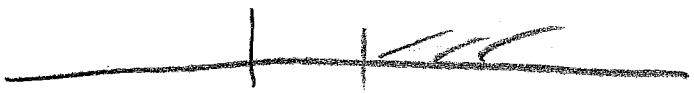
Soln $np \geq n(1-p) = 100\frac{1}{5} = 20 \geq 10$

use normal approx

Step 0) $\mu_x = np = 100\frac{4}{5} = 80$

$\sigma_x = \sqrt{np(1-p)} = \sqrt{100\underbrace{\frac{4}{5}\frac{1}{5}}_{\text{Sumstat}}} = \sqrt{16} = 4$

Step 1

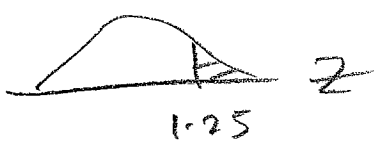


$\times \quad \mu_x = 80 \quad \sigma_x = 4$

Step 2 Standardize $z = \frac{\text{xvalue} - \mu_x}{\sigma_x} = \frac{85-80}{4}$

$= \frac{5}{4} = 1.25$

Step 3

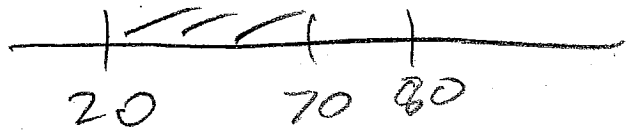


Step 4 table A $P(z \geq 1.25) = 1 - P(z \leq 1.25)$

$= 1 - .8944 = .1056$

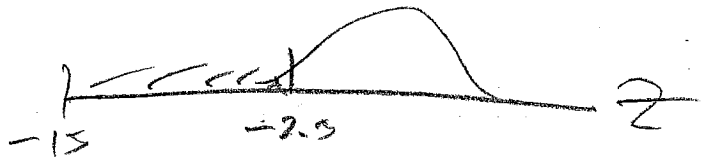
iii) Find Prob Student gets between 20 and 30 (30)
70 Questions correct.

step 0

step 1 

$X,\ \mu_X = 80 \quad \sigma_X = 4$

step 2 $\frac{70-80}{4} = -2.5$ $\frac{20-80}{4} = -15$

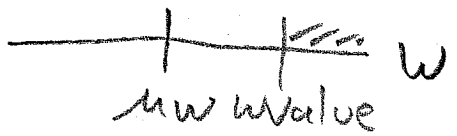
step 3 

$= P(Z \le -2.5) - P(Z < -15)$

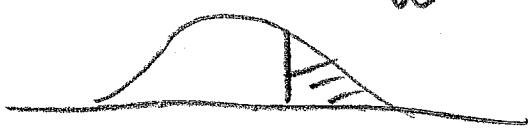
$= .0062 - 0 = .0062$

21) 3 calculations with table A ← later

step 0) Find $W \quad \mu_W \quad \sigma_W$

step 1 

step 2 Find $Z = \frac{w\ value - \mu_W}{\sigma_W}$

step 3 

step 4 table A

step 0) $\begin{cases} \bar{X} & \mu_{\bar{X}} = \mu \quad \sigma_{\bar{X}} = \sigma/\sqrt{n} \\ X & \mu_X = np \quad \sigma_X = \sqrt{np(1-p)} \end{cases}$

Key words

individual $n=1$
mean $n>1$
count of categorical

decide which to use

~~10.2.2~~ 32

§6.1 1) p298 Statistical inference assumes that data comes from a random sample or randomized experiment. If this assumption is false, your conclusions may be way off.

ex Jol resp sample Ann Landers

2) After an experiment, all probabilities are 0 or 1.

ex) Flip coin. Before the expt (flip)

$P(H) = P(T) = \frac{1}{2}$. Suppose that after the flip, the coin is H. The P(it was H) = 1

P(it was a T) = 0.

3) p306 Large sample confidence interval (CI) for μ when σ is known.

Take a SRS of size n (n large enough for CLT to hold, eg pop normal) from a pop with unknown mean μ and known SD σ

Then a 100·C% CI = $(1-\alpha)100\%$ CI for μ is $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ where $P(Z \leq z^*) = 1 - \frac{\alpha}{2}$

so $P(-z^* \leq z \leq z^*) = C$, when Z is $N(0,1)$.

4) p306 (325)

z^*	1.645	1.96	2.576
C 100%	90%	95%	99%

eg 90% CI $P(Z \le z^*) = 1 - \frac{\alpha}{2}$

$1-\alpha = .9$ so $\alpha = .1$ $\quad 1-\frac{\alpha}{2} = .95$

$P(Z \le z^*) = .95$ so $z^* = 1.645$

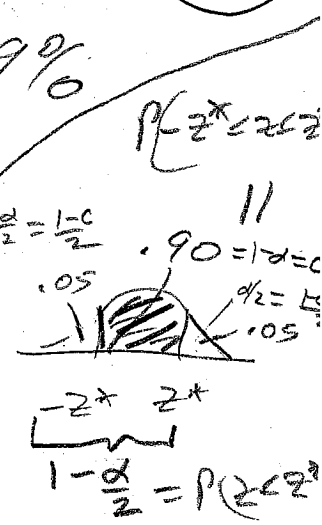


Table C row z^* has many values

ex 99% z^* 2.326

99%

5) Idea: i) $\bar{x}$ is approx $N(\mu, \frac{\sigma}{\sqrt{n}})$ if CLT holds

Let $A = P\left(\mu - \frac{2\sigma}{\sqrt{n}} \le \bar{x} \le \mu + \frac{2\sigma}{\sqrt{n}}\right) \approx .95$

Since P ($\bar{x}$ is within $2\sigma_{\bar{x}}$ of $\mu_{\bar{x}}$) $\approx .95$ by the 68-95-99.7 rule.

So $A \approx P\left(-\frac{2\sigma}{\sqrt{n}} \le \bar{x} - \mu \le \frac{2\sigma}{\sqrt{n}}\right) =$

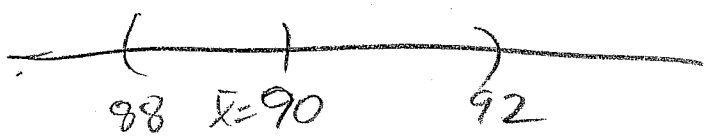
$P\left(-\bar{x} - \frac{2\sigma}{\sqrt{n}} \le -\mu \le -\bar{x} + \frac{2\sigma}{\sqrt{n}}\right)$

$= P\left(\bar{x} - \frac{2\sigma}{\sqrt{n}} \le \mu \le \bar{x} + \frac{2\sigma}{\sqrt{n}}\right) \approx .95$

ii) Suppose $\sigma = 10$, $n = 100$ $\bar{x} = 90$ and CLT holds. Then $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 1$ and $2\sigma_{\bar{x}} = 2$. Want to know for what values of μ is it "plausible" that $\bar{x}$ came from a $N(\mu, \sigma = 10)$ distribution. Fix μ and take "plausible" to mean that the chance that $\bar{x}$ is within $2\sigma_{\bar{x}}$ of μ is at least 95%.

eg $\mu_1 = 90$ $\mu_2 = 400$ $\mu_3 = 88$ $\mu_4 = 87$ $\mu_5 = 92$.

μ_3 is the smallest "plausible" value (by the 68-95-99.7 rule) and $\mu_5 = 92$ is the largest plausible value, so the 95% CI for μ is $\approx \bar{x} \pm 2 = (88, 92)$ (1.96 is better)



Plausible values of μ

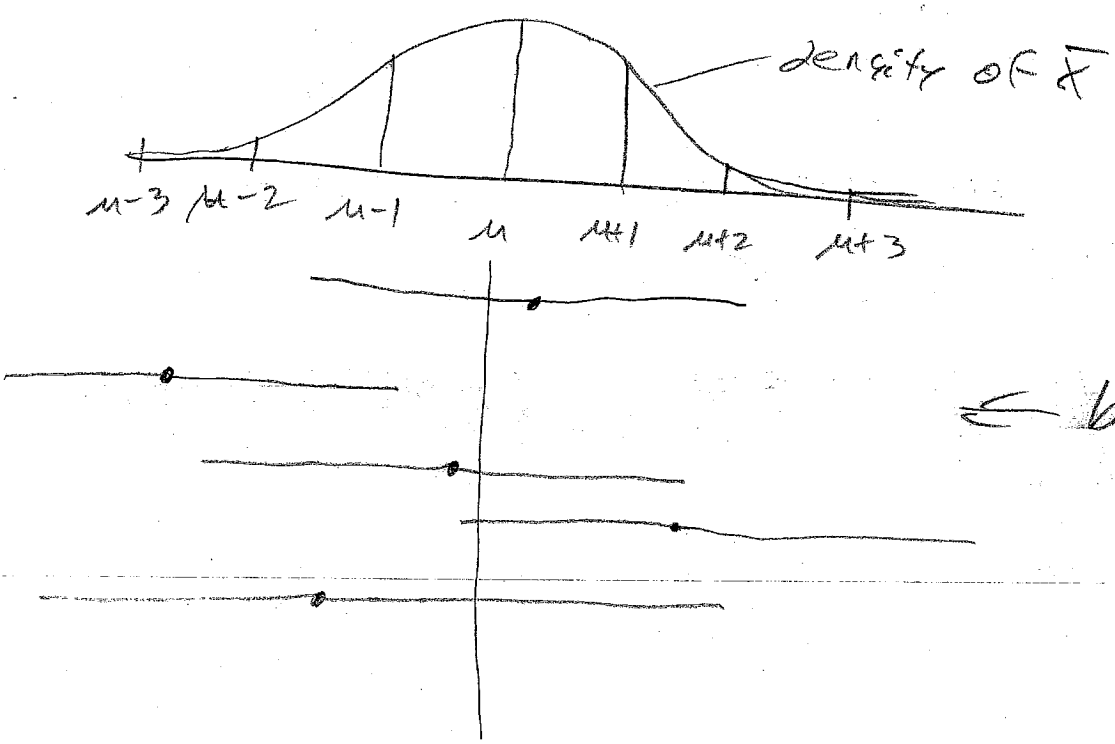
6) p301 Why 95% <u>confidence</u> instead of 95% prob?

A CI is computed <u>after</u> the experiment. Hence P(CI contains μ) = $\begin{cases} 0 & \text{if it does not} \\ 1 & \text{if it does} \end{cases}$.

If you make 100 95% CI's (100 different SRS's), approx 95 will contain μ and 5 will not.

See Figure 6.4 p 302

Suppose $\bar{x}$ is $N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = 1)$



Each CI is $(\bar{x} - 1.96, \bar{x} + 1.96)$ and has length $2(1.96) = 3.92$. If $\bar{x}$ is more than 1.96 away from μ, then the CI will NOT contain μ

If you get a "bad sample," the CI does not contain μ. Chance of bad sample = 5% for a 95% CI.

7) p308 margin of error $(= z^* \frac{\sigma}{\sqrt{n}})$ ~~.0282~~ 2/34

$= \frac{1}{2}$ CI length is a measure of how accurate we believe our estimate of the parameter (μ) is.

MBA salaries in 1994 if got MBA in 70's and are sole source of income

ex $n = 91 \quad \bar{x} = 124\,510 \quad \sigma = 180000$

Find a 90% CI for μ.

$z^* = 1.645$ eg table C

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 124510 \pm 1.645 \frac{180000}{\sqrt{91}}$$

$$= 124510 \pm 3103.97$$

$$= (121406.03,\ 127613.97)$$

8) p309 $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ CI length increases if c d
if σ doe

decreases if n increa

9) know p341 For a $c(100\%) = (1-\alpha)100\%$ CI to estimate μ within a margin of error m, take a SRS of size $n = (z^*\sigma)^2$ round up.

Ex Want to estimate cube weight to within 0.1 oz with 95% confidence, and $\sigma = 1$ oz. Find n.

Soln $m = 0.1$ $1-\alpha = .95$

$\alpha = .05$ $P(Z < 1.96) = 1 - \frac{\alpha}{2} = .975$

So $z^* = 1.96$ (or table $<$).

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = \left(\frac{(1.96)\ 1}{0.1}\right)^2 =$$

$(19.6)^2 = 384.16$

So take $n = 385$

(round up so level is at least 95%)

10) P 392 Use CI if

i) data is SRS from pop

or ii) from a randomized expts later

ii) n large enough for CLT to hold (random sample from all subpopulations of interest)

11) Sampling methods other than SRS's such as multistage samples give distributions that do not have $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

So $\bar{x} \pm 2^* \frac{\sigma}{\sqrt{n}}$ is not correct for probability methods more complex than simple random samples,

ex CI from voluntary response sample is not trustworthy

§6.2 12) p321 The null hypothesis = H_0 is a claim about a pop characteristic (parameter, such as $H_0: \mu = \mu_0$ (eg $\mu_0 = 125$ lbs)

13) The alternative hypothesis H_A is a competing claim eg H_A $\mu \begin{matrix} < \\ > \\ \neq \end{matrix} \mu_0$.

ex In ch 6 and 7, H_0 is of the form "parameter = value"

H_A " $\begin{matrix} > \\ \neq \\ < \end{matrix}$ "

Which is correct? a) $H_0: \mu > 9$ $H_A: \mu = 9$

b) $H_0: \bar{x} = 9$ $H_A: \bar{x} > 9$

c) $H_0: \mu = 9$ $H_A: \mu > 9$

14) A test statistic is used to decide whether to reject H_0 or fail to reject H_0

15) p321 The p value is the probability, assuming H_0 is true, of obtaining a test statistic at least as extreme or

"contradictory" to Ho as what was actually observed. "contradictory" is determined by the form of H_A,

Ho: the difference of the test statistic from the Ho model is due to chance

H_A: the difference of the test statistic from the Ho model is real

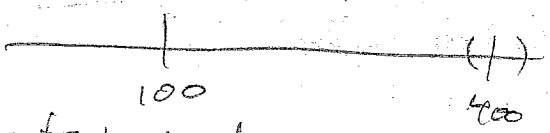
Pvalue near 0 means test statistic was very unlikely to occur if Ho was true

ex) H_0 $\mu = 100$ (σ known to be 1)

H_A $\mu > 100$

Sample of size 100 $\bar{x} = 400$

$\sigma_{\bar{x}} = \frac{1}{\sqrt{100}} = .10$



would expect difference to be real,

If $\bar{x} = 90$, Ho is less contradictory than H_A although neither seems likely,

16) A test has 4 steps p. 328 and examples

i) State Ho and H_A

ii) calculate the test statistic

iii) Find the P value.

iv) State a conclusion!

If pvalue $\leq \alpha$, reject Ho

otherwise fail to reject Ho.

Say in words what rejecting or failing to reject Ho means.

17) p327 If α is not given, use $\alpha = .05$.
Test is statistically significant if pvalue $\leq \alpha$.
Note: small pvalue is evidence against H_0

18) p329 SRS of size n, unknown μ, known σ
CLT holds if H_0 is true.

H_0 $\mu = \mu_0$ vs H_A $\mu \begin{matrix} > \\ < \\ \neq \end{matrix} \mu_0$

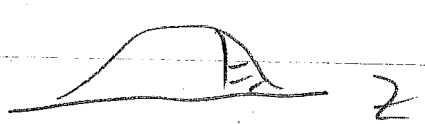
test statistic $z_0^* = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \simeq N(0,1)$ if H_0 is true

← obtained from H_0

major typo in text

pvalues

right tail test	left tail test	2 tail
H_A $\mu > \mu_0$	H_A $\mu < \mu_0$	H_0 $\mu \neq \mu_2$



pval $= P(z > z_0^*)$
$= 1 - P(z < z_0^*)$ table A

pval $= P(z < z_0^*)$

pval $= 2P(z > |z_0^*|)$
$= 2P(z < -|z_0^*|)$ table A

recall absolute value $|z| = z$, $z \geq 0$
$-z$ $z < 0$

so $|-1| = 1$, $|7| = 7$, $|-18| = 18$ etc.

19) p338 For a level α 2 tail test,
H_0 $\mu = \mu_0$ is rejected if μ_0 is outside of a $c100\% = (1-\alpha)100\%$ CI for μ,
otherwise fail to reject H_0.
(eg $\alpha = 5\%$ test is rejected if μ_0 is outside a 95% CI)

36-5

20) P343

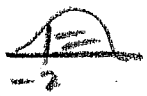
To decide the form of H_A, choose the form that goes with strong evidence.

$\bar{x} = 35 \quad \sigma = 5 \quad n = 25$

ex) Ford says Escort gets at least 37 MPG

i) $H_0: \mu = 37$ $H_A: \mu > 37$ ii) $Z_0^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{35-37}{5/\sqrt{25}} = -2$

iii) $pval = P(Z > -2) = 1 - .0228 = .9772$



iv) Fail to reject H_0 Escort does not get at least 37 MPG

ex) Consumer group says Escort gets less than 37 MPG

i) $H_0: \mu = 37$ $H_A: \mu < 37$ ii) $Z_0^* = \frac{35-37}{5/\sqrt{25}} = -2$

iii) $pval = P(Z < -2) = .0228$

iv) $pval < \alpha = .05$ so reject H_0. There is strong evidence that Escort does not get 37 MPG.

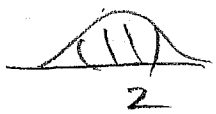
ex $H_0: \mu = 100 \quad \alpha = .05 \quad Z_0^* = 2.0$

right tail $H_A: \mu > 100$

$pval = P(Z > 2)$

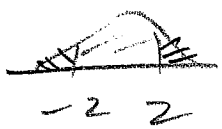
$= 1 - .9772 = .0228 < .05$ reject H_0

left tail $H_A: \mu < 100$



$pval = .9772 > .05$ fail to reject H_0

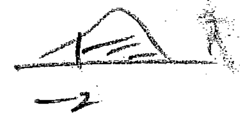
2 tail $H_A: \mu \neq 100$



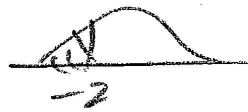
$pval = 2P(Z < -2) = 2(.0228) = .0456 < .05$

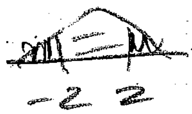
reject H_0

ex $H_0: \mu = 100 \quad \alpha = .05 \quad Z_0^* = -2.0$

right tail HA $\mu > 100$ pval =  .0282 37

$= 1 - P(z < -2) = 1 - .0228 = .9772 > .05$
fail to reject Ho

left tail HA $\mu < 100$  = Pval = $.0228 < .05$
reject Ho

2 tail HA $\mu \neq 100$  = pval = $2P(z < -2)$

~~p63~~ = $2(.0228) = .0456 < \alpha = .05$ reject Ho

21) p345 Assume ~~σ~~ (Skip p 334 – 335)
To use test need SRS
or independent observations from pop.
Need n large enough for CLT to hold
~~see ex's in book~~ I will do ex's when
both μ and σ are unknown.
(need $n \le \frac{\text{pop size}}{10}$, almost always true).

~~p32~~) 22) p343 Pvalue = .049   pvalue = .051
are not practically different. ~~Stating a pvalue~~. Always using $\alpha = .05$ makes little sense. State the pvalue and let your audience decide if the result is significant.

23) p344 Statistical significance $\neq$ practical significance.
Ho $\mu = 475$ = mean math score
$\bar{x} = 478$ 475 and 478 .. hardly different

but the pvalue could be 0.00000001 / 35
if the sample size is large enough.

24) p345 Statistical significance says that something other than chance caused a small pvalue but does not say what the cause! surgery reduced pain about so did giving patients a gel capsule placebo effect

25) p346 Suppose $\alpha = .05$. If you test many effects, about 5% will be declared significant (reject Ho) even when they are not. Alternatively, if someone unethical wants to state an effect is significant, they could take 20 samples. About 1 of them should lead to rejecting Ho even if Ho is true.

ex There is a huge on going study of nurses. Every nurse fills out a big questionnaire. Every few months some association between eg diet weight smoking and say cancer obesity etc is found and published.

The study can suggest associations, but much more work needs to be done to declare that the behavior is causing bad health. Need controlled experiments (Section 3.2

§6.4 26) p350

decision	Ho true	Ho false
reject Ho	Type I error	correct decision
fail to reject Ho	correct decision	Type II error

Type I error: reject Ho when Ho is true

Type II error: fail to reject Ho when Ho is false

27) p 350 $\alpha = P$(type I error if Ho is true).

So for $\alpha = .05$, Ho will be rejected 5% of the time when Ho is true; (100 samples about 5 will reject Ho).

Skip power p 354

Ch 7 If you like beer read p 364.

1) For ch 7 assume $n \leq \frac{\text{pop size}}{10}$ unless pop size is [illegible].

1) p 367 An SD estimated from the data is called a standard error SE.

ex) $\sigma_{\bar{x}} = SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, $SE(\bar{x}) = \frac{s}{\sqrt{n}}$.

2) p 367 If a SRS of size n is drawn from a $N(\mu, \sigma)$ population, then the sampling distribution of

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ has a <u>t distribution with</u> n-1 degrees of freedom, denoted by $t(n-1)$.

B) p370 Suppose a SRS of size n is drawn from a pop with unknown mean μ and SD σ. If $n \geq 2$ is large enough for the CLT to hold, then a $100(1-\alpha)\% = 100c\%$ CI for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} \quad \text{where}$$

$P(t_{(n-1)} \leq t^*) = 1 - \frac{\alpha}{2}$, or $P(-t^* \leq t_{(n-1)} \leq t^*) = c$

4) Get t^* from table C. If $df = n-1 > 30$, use z^* rather than df 40-1000.

ex] $\bar{x} = 100 \quad n = 16 \quad s = 40$

$df = 16 - 1 = 15$ Find a 95% CI assume CLT holds

df	2.131
15	95% conf level C

95% CI is $100 \pm 2.131 \frac{40}{\sqrt{16}} = 100 \pm 21.31$

$= (78.69, 121.31)$

ex] $\bar{x} = 100 \quad n = 32 \quad s = 40$ Find 95% CI

$df = 31 > 30$ go go to $z^* = 1.96$

$100 \pm 1.96 \frac{40}{\sqrt{32}} = 100 \pm 13.859 = (86.141, 113.859)$

Using df = 30 would give 100 ± [illegible]

But rather than interpolating or taking closest, just use z*. A t(n-1) distn is almost a N(0,1) dist for n-1 ≥ 30 or n ≥ 31

i.i.d key assumption is n large enough for CLT to hold, eg pop is normal

5) p 370 One sample t-test

SRS of size n ≥ 2 large enough for CLT to hold from a pop with unknown SD σ;

Test $H_0: \mu = \mu_0$

test stat $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \approx \begin{cases} t(n-1), & n-1 \le 30 \\ N(0,1), & n-1 > 30 \end{cases}$

use table A if $n-1 > 30$ or $n > 31$

C if $n-1 \le 30$ or $n \le 31$ to find pvalues.

left tail	right tail	2 tail
$H_A: \mu < \mu_0$	$H_A: \mu > \mu_0$	$H_A: \mu \neq \mu_0$
		
$t(n-1)$	$t(n-1)$	$t(n-1)$
t_0	t_0	$-\lvert t_0 \rvert$ $\lvert t_0 \rvert$
$n-1 \le 30$	$n-1 \le 30$	$n-1 \le 30$
$pval = P(t(n-1) \le t_0)$	$pval = P(t(n-1) \ge t_0)$	$P(val) = 2P(t(n-1) \ge \lvert t_0 \rvert)$
$= 1 - P(t(n-1) > t_0)$	$= 1 - P(\text{table}, n-1 > 30)$	$= 2P(\text{table}, n-1 > 30)$
$P(Z \le t_0)$ $n-1 > 30$		

same 4 step procedure i) State H_0 and H_A

ii) calculate test statistic t_0

iii) Find Pvalue

iv) state a conclusion

if pval $< \alpha$ reject H_0

otherwise fail to reject H_0

use $\alpha = .05$ if not given,

state in words what rejecting or failing to reject H_0 means,

ex) $H_0: \mu = 0$ $H_a: \mu > 0$ $n = 15$

normal pop $t_0 = 1.97$

a) Find df $df = n - 1 = 14$

b) Find the 2 t^* values from table that bracket t_0.

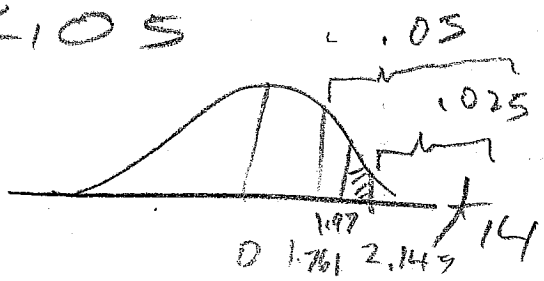
df	upper tail prob .05	.025
14	1.761	2.145

c) what are the right tail probs of these 2 t values?

.05 and .025

d) What is the p value?

$.025 < \text{pvalue} < .05$



e) Is the test

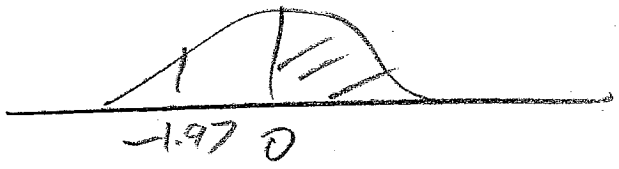
significant at 5% level?

1% level?

pval < 5% so significant at 5%

pval > 1% so not sig at 1%.

i) if $t_0 = -1.97$, what is pval .0282



pval > .05 fail to reject H_0

in fact $.95 < \text{pval} < .975$

$1-.05$ $1-.025$

ii) $t_0 = 1.97$ H_A $\mu < 0$ what is pval

pval > .5

in fact $.95 < \text{pval} < .975$, fail to reject H_0

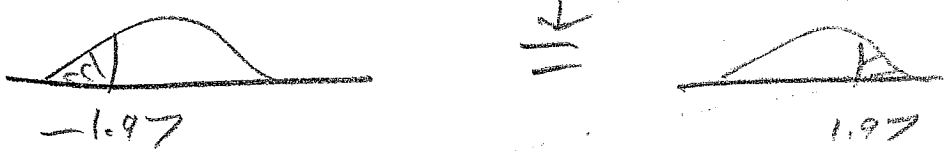
iii) $t_0 = 1.97$ H_A $\mu \neq 0$

$2(.025) < \text{pval} < 2(.05)$

or $.05 < \text{pval} < .1$ fail to reject H_0

iv) $t_0 = -1.97$ H_A $\mu < 0$

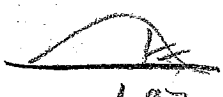
by symmetry



so $.025 < \text{pval} < .05$

v) $t_0 = 1.97$ $n = 32$ H_A $\mu > 0$.

$df = 31 > 30$ so use table A



$= 1 - .9756 = 0.0244$

ex) $\alpha = .01$ Golpher tries out club to see if new club drives balls further than old club. with old club he believes he can hit ball 200 yards. Will buy club if it is better. Shop lets him take 10 drives. $n = 10$ $\bar{x} = 204.60$ $s = 9.03$

soln μ = mean drive length with new club

step 1) $H_0: \mu = 200$ μ_0 is never from sample
$H_A: \mu > 200$

step 2) $t_0^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{204.60 - 200}{9.03/\sqrt{10}} =$

$\frac{4.6}{2.8555} = 1.611$

step 3) $df = n-1 = 9$ $pval = P(t_9 > 1.61)$

= 

table c

df	.10	.05
9	1.383	1.833

$.05 < pval < .1$

step 4) Fail to reject H_0 since $pval > \alpha = .01$.

Not enough evidence to conclude that the new club is better than the old club.

Common final problem

Test from computer output — 0282 — 45

amount of relief from sugar pill

Ex Test of $\mu = 0.000$ vs μ N.E. 0.00

	N	MEAN	STDEV	SEMEAN	T	P
sugar pill	8	4.000	4.3094	1.5236	2.63	0.0341

Step 1) H_0 $\mu = 0$ H_A $\mu \neq 0$

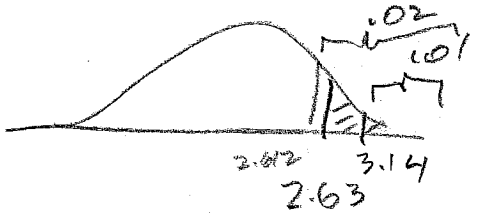
Step 2) $t_0^* = 2.63$

Step 3) pvalue = .0341

Step 4) pval = .0341 < .05 so reject H_0

Sugar pill brings some relief

Note computer pvalue is better than values in table C.



df	.02	.01
7	2.612	3.143

2 tail test so

$2(.01) = .02 \leq$ pval $\leq .04 =$

7) Matched Pairs) + IP375 - U

Data comes in pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

x_i and y_i are dependent, that is knowledge of x_1 should give some information about y_1. The pairs (x_i, y_i) and (x_j, y_j) are independent for $i \neq j$.

That is knowledge of the 1st pair gives no info about the 2nd pair etc. uncorrelated

Often 2 measurements are taken from the same subject

eg X_i = cholesterol level for ith person before med
Y_i after

correlated so use matched pairs

ex] age of husband age of wife

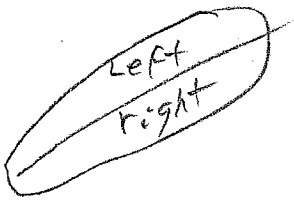
compute $D_1, \ldots, D_n$ where $D_i = X_i - Y_i$
to get 1 sample of D's.

A) Perform a t CI on the D's

B) perform a t test on the D's

ex] Matched paired t test

Use 2 different viral preparations on a tobacco leaf. Divide leaf into left half right. flip coin, put virus 1 on left if heads otherwise on.



Count number of lesions due to each
Dependent since strong leaf should resist both viruses weak plant or will have less resistance.

Suppose researchers expect virus 1 to cause more lesions than virus 2.

$\bar{D} = \frac{\sum D_i}{n} = \frac{\sum (X_i - Y_i)}{n} = 1.450$ $S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}} =$

$n = 20$

μ_D = mean of $X_i - Y_i$

4 step test

$x_i > y_i$ so $x_i - y_i > 0$ 0282 4/2

Step 1 $H_0: \mu_D = 0$ $H_A: \mu_D > 0$

Step 2 $t_0 = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} = \frac{1.450 - 0}{3.203/\sqrt{20}} = 2.02$

Step 3 $df = 19$ table C

	.05	.025
19	1.729	2.1093

So $.025 < pval < .05$

Step 4) reject H_0 since $pval \leq .05$.

Virus 1 causes more lesions than Virus 2.

Find a 90% CI

$\bar{D} = 1.450$ $s_D = 3.203$ $n = 20$ $df = 19$

table C $t^* = 1.729$

90% CI for μ_D is

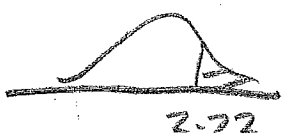
$1.450 \pm 1.729 \frac{3.203}{\sqrt{20}} = 1.450 \pm 1.238 = (0.212, 2.6$

ex) Suppose $n = 36$ instead of 20

Step 1 $H_0: \mu_D = 0$ $H_A: \mu_D > 0$

Step 2 $t_0 = \frac{\bar{D} - \mu_D}{s_D/\sqrt{36}} = \frac{1.450 - 0}{3.203/\sqrt{36}} = \frac{1.450}{0.5338} =$

Step 3 $df = 36 - 1 = 35 > 30$ so use table A.



$= 1 - .9967 = 0.0033$

(table C w
Say .002
pval <
table A is more

Step 4. pval $< \alpha = .05$ so reject H_0

Viral produces more lesions than Virus

ex) p377 output) i) $H_0: \mu_D = 0$ $H_A: \mu_D > 0$ ii) $T^* = .35$ iii) pval = .365 iv) fail to reject at H_0

8) p380 When to use t test tCI

SRS from population (with finite SD σ)
or data measurements from an experiment

a) $n < 15$ Use if data is close to normal.
If data is skewed or outliers are present
don't use t procedures.

b) $n \geq 15$ Use t test unless outliers are present or the data is strongly skew

c) If n is large enough, t procedures can be used even on strongly skewed data.

Text suggests $n \geq 40$ for clearly skewed data

d) Don't use a t procedure if the pop mean is known

See ex 7.5 p380

ϕ 7.2
p390

9) 2 sample procedures (other than matched pairs) need 2 independent samples
or an experiment that used randomization to form a treatment and control group (1/2 get medic 1/2 get plac

9.2 p392 | site | mean | SD | 0282

10) * SRS n_1 $\bar{x}_1$ s_1

SRS n_2 $\bar{x}_2$ s_2

p395

df = k = smaller of $n_1 - 1$, $n_2 - 1$

if output is not being used,

H_0 $\mu_1 = \mu_2$ or H_0 $\mu_1 - \mu_2 = 0$

(or H_0 $\mu_1 - \mu_2 = a$).

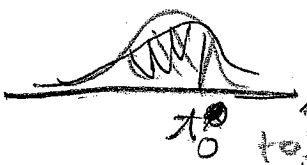
test statistic $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

p values

right-tail	left-tail	2-tail
H_A $\mu_1 > \mu_2$	H_A $\mu_1 < \mu_2$	H_A $\mu_1 \neq \mu_2$

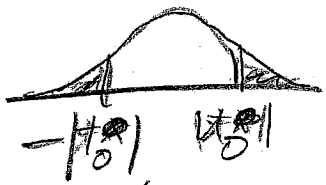
$P(t_k > t_0^*)$ $k \leq 30$

$1 - P(z \leq t_0)$, $k > 30$ p395



$P(t_k < t_0^*)$ $k \leq 30$

$P(z < t_0)$ $k > 30$ table A



$2P(t_k > |t_0^*|)$,

$2P(z < -|t_0|)$ k

12) * 100C% = 100(1−α)% CI for $\mu_1 - \mu_2$

is $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

where $P(-t^* < t_k < t^*) = C = 1 - \alpha$

or $P(t_k < t^*) = 1 - \alpha/2$, $k = df$ = smaller of $n_1 - 1, n_2$

12) *p405 Use 2 sample t if $\mu_1, \mu_2, \sigma_1, \sigma_2$ are unknown, no reason to believe that $\sigma_1 = \sigma_2$.

pop: x_1, x_2 — mean: μ_1, μ_2 — SD: σ_1, σ_2

If $n_1 = n_2$ and the shapes of the 2 histograms are similar, can use t test even for $n_1 = n_2$ small! ($n_1 = n_2 \geq 5$ text+)

If $n_1 \neq n_2$ want CLT to hold for both samples.

13) p405 ~~pooled~~ now omitted 2 sample t tests and CI's are used if $\mu_1, \mu_2, \sigma_1, \sigma_2$ are unknown and it is believed (before the experiment) that $\sigma_1 = \sigma_2$.

You only need to know how to do the test from output.

ex p405 — SAS output

	variances	T	DF	Prob > |T|
assume SD's are equal	Equal	2.9912	10.0	0.0135

step 1 H_0 $\mu_1 = \mu_2$ H_A $\mu_1 \neq \mu_2$

step 2 $T_0 = 2.9912$

step 3 pval = .0135

step 4) ~~pval = .0135~~ pval = .0135 < α = .05 reject H_0 table C 2(.005) < pval or .01 < pval < .02

DDT means (space height) not equal for control and mean c.o.c.

df	.01	.0
10	2.764	3.16

Warning Do not look at sample variances to decide whether to use 2 sample t or pooled 2 sample t. Use pooled 2 sample t if the story problem says the population SD's or variances are approx. equal.

Legal, 7.47

ex) Bank has 2 proposals to increase the amount of credit charged on its credit cards. Each proposal is offered to a SRS of 150 of its customers.

proposal	n	$\bar{x}$	s
A	150	1987	392
B	150	2056	413

$\bar{x}_A$ estimates μ_A = mean (total) annual amount charged by customers who used proposal A

Test whether there is a difference in μ_A and μ_B.

Step 1 $H_0: \mu_A - \mu_B = 0$ $H_A: \mu_A - \mu_B \neq 0$

2) $t_0^* = \dfrac{\bar{x}_A - \bar{x}_B}{\sqrt{\dfrac{s_A^2}{n_1} + \dfrac{s_B^2}{n_2}}} = \dfrac{1987 - 2056}{\sqrt{\dfrac{(392)^2}{150} + \dfrac{(413)^2}{150}}} =$

$$\frac{-69}{\sqrt{1024.43 + 1137.13}} = \frac{-69}{\sqrt{2161.55}} = -1.484$$

Step 3 $df = n - 1 = 149$ so use table A

$= pval = 2\,P(z < -1.48)$

$= 2(.0694) = .1388$

iv) pval > .05 fail to reject H_0.

The ~~mean~~ charge ~~amounts~~ from ~~the~~ amounts ~~proposals~~ are ~~the same~~ by chance.
see ex 7.9 p 398-404

ex Is the angular velocity of the knee higher for skilled rowers than for novice rowers?

Output	N	mean	STD DEV	Std error	T	DF	P
Skilled	10	4.1828	0.47906	0.15149			
novice	8	3.0100	0.95895	0.3390			

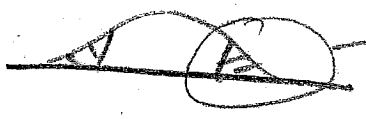
Variances	T	DF	Prob > |T|
unequal	3.1583	9.6	0.0104
equal	3.3918	16.0	0.0037

Step 1 H_0 $\mu_S - \mu_N = 0$ H_A $\mu_S - \mu_N > 0$

Step 2 $T_0 = 3.1583$

Step 3 $P\text{-val} = \frac{1}{2}(0.0104) = .0052$

SAS $p\left(t_{150} > 1.1\right)$



or check $df = 9.6 \approx 10$

3.158 ≈ 3.169

Step 4) pval < .05 so reject H_0

Skilled rowers have higher knee velocities.

ex) Same but suppose we are told that the pop SD's of the 2 groups are approx equal. Then use pooled 2 sample T.

Step 1 H_0 $\mu_S - \mu_N = 0$ H_A $\mu_S - \mu_N > 0$

Step 2 $T_0 = 3.318$ Skip

Step 3 $pval = \frac{.0037}{2} = .00185$

(from table C $.001 < pval < .0025$)

Step 4 reject H_0

skilled rowers have higher knee velocities

Skip 9.403, 9.7.3

Ch 8 CI's and Tests for proportions

1) Recall $\hat{p} = \frac{\text{count of successes}}{n}$ where n = sample size

2) *p 437

A $100(1-\alpha)\% = c\%$ CI for p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{where}$$

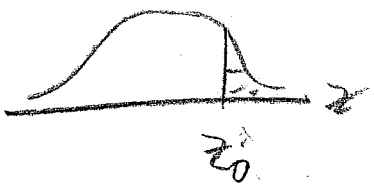
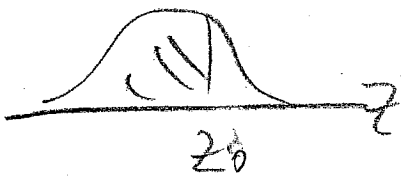
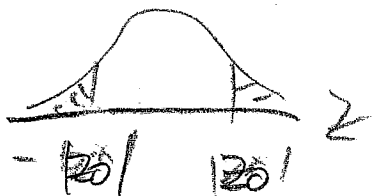
$P(-z^* \le z \le z^*) = 1-\alpha = c$ and

$P(z \le z^*) = 1 - \frac{\alpha}{2}$.

3) *p 437 test $H_0: p = p_0$

Statistic $z_0^* = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$ if H_0 is true.

Pvalues

right tail	left tail	2 tail
$H_A: p > p_0$	$H_A: p < p_0$	$H_A: p \neq p_0$
		
z_0^*	z_0^*	$-\lvert z_0^* \rvert$ $\lvert z_0^* \rvert$
$P(z > z_0^*)$	$P(z < z_0^*)$	$2P(z < -\lvert z_0 \rvert)$
$1 - P(z < z_0)$		$= 2P(z > \lvert z_0 \rvert)$

p 435

4) *Assumptions

a) Data are SRS from population.

b) Pop is at least 10 times the sample size (almost always true)

c) For test $np_0 \ge 10$ and $n(1-p_0) \ge 10$

For CI $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

6) p431-432 (similar to p278 $\hat{p} = \frac{x}{n}$)

Sampling distribution of $\hat{p}$

If $np \geq 10$ $n(1-p) \geq 10$ and $\hat{p}$ is from a SRS of size $n \leq \frac{\text{pop size}}{10}$

then $\hat{p} \approx N\left(\mu_{\hat{p}} = p,\ \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$

ex) Forwards calculation for $\hat{p}$

Suppose a politician decides to run for office if 40% or more of a SRS of 2500 voters say they would vote for her. If the true pop percentage was 38%, what is the chance the politician will run?

Soln step 1)

Step 0) $\mu_{\hat{p}} = p = .38$ (always from pop never from sample)

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.38(1-.38)}{2500}} = \sqrt{.0000942}$$

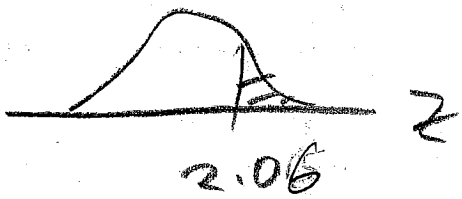
$= .00971$ step 1)  $\hat{p}$ $\mu_{\hat{p}} = .38$ $\sigma_{\hat{p}} = .00971$

step 2) get z score (triplet $\hat{p}$, $\mu_{\hat{p}}$, $\sigma_{\hat{p}}$)

$$z = \frac{\hat{p}\text{ value} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

$= \frac{.40 - .38}{.009771} = 2.05$

Step 3) z picture



Step 4 table A $1 - \underset{2.06}{(1)} = 1 - .9803$

$= .0197$

That is there is about a 2% chance she will run.

---

on p 278 we had X = count

$X \approx N(\mu_X = np \quad \sigma_X = \sqrt{np(1-p)})$

Note $\hat{p} = \frac{X}{n}$ $\mu_{\hat{p}} = \frac{\mu_X}{n} = p$ $\sigma_{\hat{p}} = \frac{\sigma_X}{n} = \sqrt{\frac{p(1-p)}{n}}$

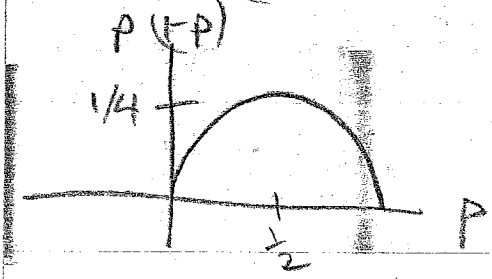
For CI $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

6) p448 Sample size for $\hat{p}$ to be within m of p with $100(1-\alpha)\%$ $= 100c\%$ confidence if

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) \quad \text{where}$$

p^* is a good guess for p and

$P(z < z^*) = 1 - \frac{\alpha}{2}$ or $P(-z^* < z < z^*) = 1-\alpha$



If no good guess for p is available, use $p^* = \frac{1}{2}$ (and $n = \left(\frac{z^*}{2m}\right)^2$).

Round up to make n an integer

(If p is far from 0.5, then using $p^* = \frac{1}{2}$ will result in using a much larger sample size than actually needed.)

ex) SRS of 1711 from records of people who died in bicycle crashes between 1987 and 1991 386 of the 1711 had blood alcohol levels above 0.10% (were drunk).

a) Find a 95% CI for p.

$$\hat{p} = \frac{386}{1711} = 0.2256$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.2256\,(1-.2256)}{1711}} = 0.0101$$

$$z^* = 1.96$$

$$95\%\,CI = \hat{p} \pm z^*\, SE(\hat{p})$$

$$= .2256 \pm 1.96\,(.0101)$$

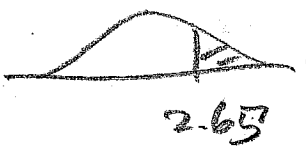
$$= .2256 \pm .0198 = (.2058,\ .2454)$$

b) Test whether p is greater than 0.20. (same 4 step procedure)

i) H_0 $p = .20$ $\quad H_A$ $p > .20$

ii) $$z_0^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.2256 - .2}{\sqrt{\frac{.2\,(1-.2)}{1711}}} = \frac{.0256}{.0097}$$

$$= 2.65$$

iii) 

$= \text{pval} = 1 - .9960 = 0.004$

iv) Reject H_0 since pval $= .004 \leq \alpha = .05$.

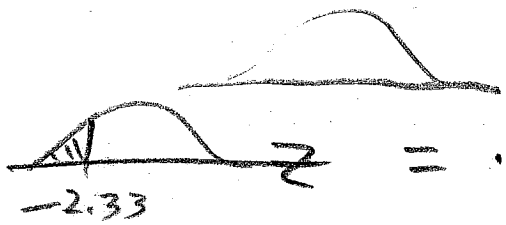
The proportion of fatally injured cyclists that were drunk > 0.20.

c) Test whether p is less than .25. since <

i) H_0 $p = .25$ vs H_A $p < .25$

ii) $z_0^* = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{.2256 - .25}{\sqrt{\dfrac{.25(.75)}{1711}}} = \dfrac{-0.0244}{.0105}$

$= -2.33$

iii) Pvalue = 

$= .0099$

iv) Reject H_0 since pvalue $\leq .05 = \alpha$,
the proportion of fatally injured cyclists that were drunk < 0.25.

(Big sample so $\hat{p} = .2256$ is fairly accurate)

D) Find n to estimate p to within 0.01 with 90% confidence if no good guess for p is available.

$n = \left(\dfrac{z^*}{m}\right)^2 \tfrac{1}{2}\left(1-\tfrac{1}{2}\right) = \left(\dfrac{1.645}{.01}\right)^2 \tfrac{1}{2}\tfrac{1}{2} = (164.5)^2 \tfrac{1}{4}$

$= \dfrac{27060.25}{4} = 6765.06$ use $n = 6766$

E) Find n if $p^* = .20$

$n = \left(\dfrac{1.645}{.01}\right)^2 .2(1-.2) = 4329.64$

use $n = 4330$

7) 2 proportions p447 SRS pop p_i n_i $\hat{p}_i$ (48

independent 1 p_1 n_1 $\hat{p}_1$

2 p_2 n_2 $\hat{p}_2$

8) p449 100(1−α)% CI for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE_{(\hat{p}_1 - \hat{p}_2)} =$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $P(z < z^*) = 1 - \frac{\alpha}{2}$ or $P(-z^* < z < z^*) = C = 1-$

9) p453 test $H_0: p_1 = p_2$

Let $\hat{p}$ = count of "successes" in both samples

= pooled sample proportion, $n_1 + n_2$

Test statistic $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$ under H_0

P vals

right tail	left tail	2 tail
H_A $p_1 > p_2$	H_A $p_1 < p_2$	H_A $p_1 \neq p_2$
		
$P(z > z^*) =$	$P(z < z^*)$	$-\lvert z^* \rvert \quad \lvert z^* \rvert$
$1 - P(z \leq z^*)$		$2P(z > \lvert z^* \rvert)$
		$= 2P(z < -\lvert z^* \rvert)$

10) Assumptions for 2 sample proportions procedures

a) 2 independent SRS's
or trt group vs control group (or another trt) created with random chance mechanism

b) $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$
$n_2\hat{p}_2 \geq 5$, $n_2(1-\hat{p}_2) \geq 5$

c) pop 1 size $\geq 10\, n_1$
pop 2 size $\geq 10\, n_2$
} almost always true

ex

		# who say women are safest driver	
100	men	31	$\hat{p}_1 = .31$
100	women	67	$\hat{p}_2 = .67$

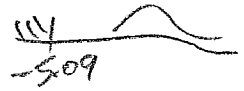
$$\hat{p} = \frac{31+67}{200} = 0.49$$

(Source www.usatoday.com snapshots who is the safest driver)

A) Test i) H_0 $p_1 = p_2$ vs H_A $p_1 < p_2$

ii) $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{.31-.67}{\sqrt{.49(.51)\left(\frac{1}{100}+\frac{1}{100}\right)}}$$

$$= \frac{-0.36}{\sqrt{.00499}} = \frac{-.36}{.0707} = -5.09$$

iii)  pval = 0.0
-5.09

iv) reject H_0 there is strong evidence that the

proportion of men who say that women are safer drivers than men is smaller than the proportion of women who say so.

B) Find a 90% CI for $p_1 - p_2$

$z^* = 1.645$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= \sqrt{\frac{(.31)(.69)}{100} + \frac{(.67)(.33)}{100}}$$

$$= \sqrt{.002139 + .002211} = \sqrt{.00435}$$

$$= 0.06595$$

90% CI is $(\hat{p}_1 - \hat{p}_2) \pm z^* SE$

$$= (.31 - .67) \pm 1.645(.06595) = -0.36 \pm 0.108$$

$$= (-.468, -.252)$$

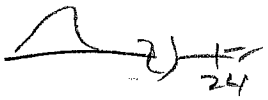
Ch 9 For hypotheses tests so far, we get a statistic $\bar{x}$, $\hat{p}$, $\bar{x}_1 - \bar{x}_2$, or $\hat{p}_1 - \hat{p}_2$ and find the appropriate z value (if df > 30). We reject H_0 if "z" is far from zero in the direction of H_A. If "z" is small or far away from zero in the direction away from H_A we fail to reject H_0.

ex: μ = mean IQ of M282 students

H_0 $\mu = 60$ H_A $\mu > 60$ $\bar{x} = 120$, $s = 15$

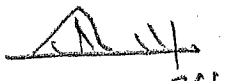
$$z \approx \frac{120 - 60}{15/\sqrt{36}} = 24.0$$

mean IQ is higher than 60



pval = 0
reject H_0

ex H_A $\mu < 60$



pval = 1

no evidence that mean IQ of 282 students is less than 60

In this ex, H_0 is absurd, but H_A is more absurd

pval = 1 means $\bar{x}$ could not have occurred if H_A is true

H_0

p.1 on review

0

1) p483 Suppose there are 2 categorical variables and we want to test

H_0 there is no relationship between the 2 categorical variables

H_A there is a relationship

2) p472 The 2 cat. variables can be displayed in a (two way) $r \times c$ table.

"# rows = columns

The row variable has r categories and the column variable has c categories

3) p472 Each entry of the table is a cell

There are rc cells. (multiply)

4) p473 A row marginal total is the sum of the row (or column)

A column

The table total is the sum of all of the counts $= n$

ex1

	gender		
	F	M	row tot
civil eng	10	20	30
r nuclear eng	30	40	70
industrial eng	40	60	100
col tot	80	120	200 = n

c

3 × 2 table
6 cells

5) p 483 The test of H_0 vs H_A can be used for 2 situations

A) There are r independent SRS's from r populations. Each person is classified into one category of column categorical variable.

B) There is a single SRS of individuals. Each individual is classified according to both of two categorical variables.

ex) In the last ex, take a SRS of 30 civil eng's 70 nuc eng's 100 ind eng's and classify each individual by gender This gives situation A.
For situation B, take a SRS of 200 civil nuclear and ind engineers. Classify each individual by type of engineer and by gender.

6) In Situation A, r SRS's, suppose the categorical variable has only two values success, failure so $c = 2$

eg ~~male~~ female, count females

eg cured not cured count cured

Then the sample proportion from each SRS can be obtained and plotted in a bar graph. Consider

H_0 there is no relationship between the two categorical variables

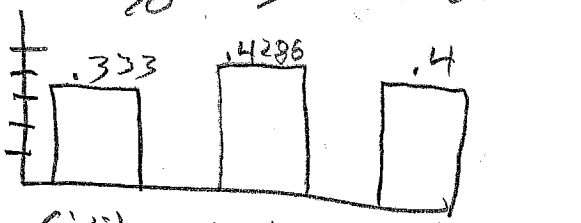
H_A there is a relationship

If H_0 is true then there is no difference between the r SRS's ("treatments") and $p_1 = p_2 = \cdots p_r$. If H_0 is not true then there is a difference so not all of the $p_1, p_2, \ldots, p_r$ are equal.

ex eng vs gender $\hat{p}_i$ = prop of women from ith major

$\hat{p}_1 = \frac{10}{30} = \frac{1}{3} \approx .333$ $\hat{p}_2 = \frac{30}{70} = \frac{3}{7} \approx .4286$

$\hat{p}_3 = \frac{40}{100} = .40$



Is there a difference in bar chart heights or is the difference due to chance?

7) p 472 If H_0 is true, then

$$\text{the expected cell count} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

8) p476 The test statistic for Ho is the chi-square statistic

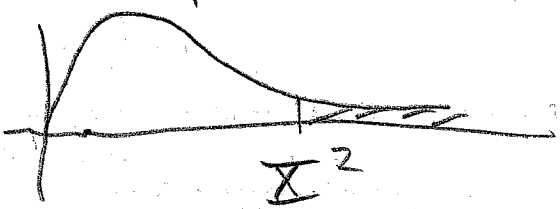
$$\chi^2 = \sum_{\text{sum over rc cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

r = # rows
c = # columns

Cell components of χ^2 p478

9) p480 If Ho is true $\chi^2 \approx \chi^2_{(r-1)(c-1)}$ df

and the pvalue is from table E



$pval = P(\chi^2_{(r-1)(c-1)} \geq \chi^2)$

If pvalue $\leq \alpha$, reject Ho
conclude there is a relationship.

Same 4 step test.

10) p485 use the χ^2 test if no more than 20% of the expected counts are less than 5 and if all of the individual expected counts are 1 or greater. (Also see p478.)

Ex A 2x2 table has 4 cells and .2(4) = .8 < 1. So all 4 expected cell counts should be 5 or greater.

ex eng

	Gender F	M	
Fr	80	120	200
So	80	120	200
Jr	60	90	150
Sr	80	90	150

proportions by major

.4	.6
.4	.6
.4	.6
.4	.6

So knowing row proportions tells nothing about year, hence gender and year are unrelated.

ex

	uses nail polish	does not use nail polish	row prop's
M	0	200	0.0 1.0
F	120	80	1.6 .4

now there is a relationship

Note Expected count for a cell ij in ith row jth column is $\frac{\text{ith row total} \cdot \text{jth column total}}{n}$

$= n \left(\frac{\text{ith row total}}{n}\right)\left(\frac{\text{jth col tot}}{n}\right) \approx n\,(\text{prob in ith row})(\text{prob in jth col})$

$= n\,P$ (in ith row and in jth column)

if col and row are independent (mult rule)

ex p472 – 4>4

	relapse: no	relapse: yes	row total
Desip	14	10	24
Lith	6	18	24
placebo	4	20	24
col tot	24	48	72

Test: i) H_0 no relationship between relapse & treatment

H_A there is a relationship

ii) work table of obs

exp [ ] = $\frac{\text{row tot} \cdot \text{col tot}}{\text{table tot}}$

χ^2 contrib ( ) = $\frac{(O-E)^2}{E}$

			row tot
	14	10	
	[8]	[16]	24
	(4.5)	(2.25)	
	6	18	
	[8]	[16]	24
	(0.5)	(0.25)	
	4	20	
	[8]	[16]	24
	(2.0)	(1.0)	
col tot	24	48	72

$\frac{24 \cdot 24}{72}$ obs; exp; cell χ^2; $\frac{(14-8)^2}{8}$; $\frac{(10-16)^2}{16}$; $\frac{(6-8)^2}{8}$; $\frac{(18-16)^2}{16}$; $\frac{(4-8)^2}{8}$; $\frac{(20-16)^2}{16}$; $\frac{24(48)}{72} = 16$

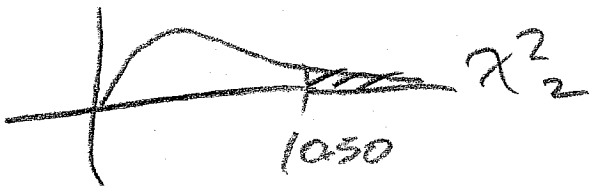
χ^2 = sum of χ^2 cell contributions

$= 4.5 + 2.25 + 0.5 + 0.25 + 2.0 + 1.0 = 10.50$

iii) $df = (r-1)(c-1) = (3-1)(2-1) = 2$

table E

df	.01	.005
2	9.21	10.60



$.005 < pval < .01$

iv) reject Ho there is a relationship between treatment and relapse

ex) Minitab output makes test much easier
I am likely to give table with χ^2 cell contributions added, but I will leave some expected counts and χ^2 cell contributions blank.

---

p477

HCY line chisq = ...

= 10.5

df = 2 pval = .005

i) Ho there is no relationship between trt and relapse
Ha a relationship

ii) $\chi^2 = 10.5$

iii) pval = .005

iv) reject Ho there is a relationship between trt and relapse

ex SRS 125 college men 5765
150

		days per week exercising (women)				
		0-1	2-3	4-5	6-7	row tot
M	obs	40	53	26	6	125
	exp	33.636	55.0	28.636	7.727	
	cell chisq	1.204	.073	.243	.386	
F	obs	34	68	37	11	150
	exp	40.363	66.0	34.364	9.273	
	cell chisq	1.003	.061	.202	.322	
	col tot	74	121	63	17	275

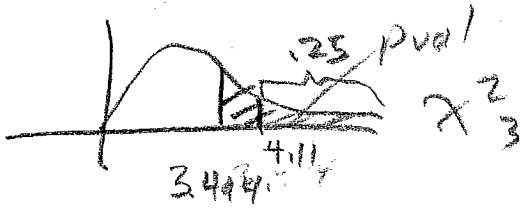
4 step test $\chi^2 = 1.204 + .073 + .243 + .386 + 1.003 + .061 + .202 + .322 = 3.494$

i) Ho there is no relationship between gender and # days exercising
HA " " a "

iv) $\chi^2 = 3.494$

iii) $df = (r-1)(c-1) = (2-1)(4-1) = 3$

df	table E .25	.20
3	4.11	4.61



$.25 \leq pval \leq 1$

v) Fail to reject Ho
There is no relationship between gender and # of days exercising

ex SRS of 2453 fatally injured pedestrians

age	blood alcohol conc	0.0	0.01 – 0.09	0.10 –	tot
16–34	obs	439	85	696	1220
	exp	473.477	91.015	655.508	
	cell chisq	2.510	0.398	2.501	
35 –	obs	513	98	622	1233
	exp	478.523	91.985	662.492	
	cell chisq	2.484	0.393	2.475	
tot		952	183	1318	2453

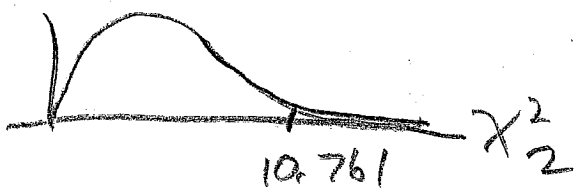
$\chi^2 = 2.510 + 0.398 + 2.501 + 2.484 + 0.393 + 2.475 = 10.761$

i) H_0 age at death and blood alcohol conc are not related

H_A are related

ii) $\chi^2 = 10.761$

iii) $df = (2-1)(3-1) = 2$



df	.005	.0025
2	10.60	11.98

$.0025 < pval < .005$

iv) Reject H_0 age at death and blood alcohol conc are related.

FM 285 S45

11) A chisquare test can also be used for a SRS from a single population where each individual is classified according to both of two categorical variables.

4 step test

i) Ho: no relationship between the 2 categorical variables

H_A: there is a relationship

ii) $X^2 = \sum \frac{(obs - exp)^2}{exp} \approx \chi^2_{(r-1)(c-1)}$

sum over the rc cells.

iii) Pvalue $= P(\chi^2_{(r-1)(c-1)} > X^2)$

$\chi^2_{(r-1)(c-1)}$ table E

iv) If pval $\leq \alpha$, reject Ho

there is a relationship.

12) p 544

Use χ^2 test if no more than 20% of expected counts are less than 5 and all individual expected counts are 1 or greater.

ex: ... 12 (4) = .48 < 1. so all 4 ...

Ch 10 1) p531-532 recall regression $\hat{y} = a + bx$

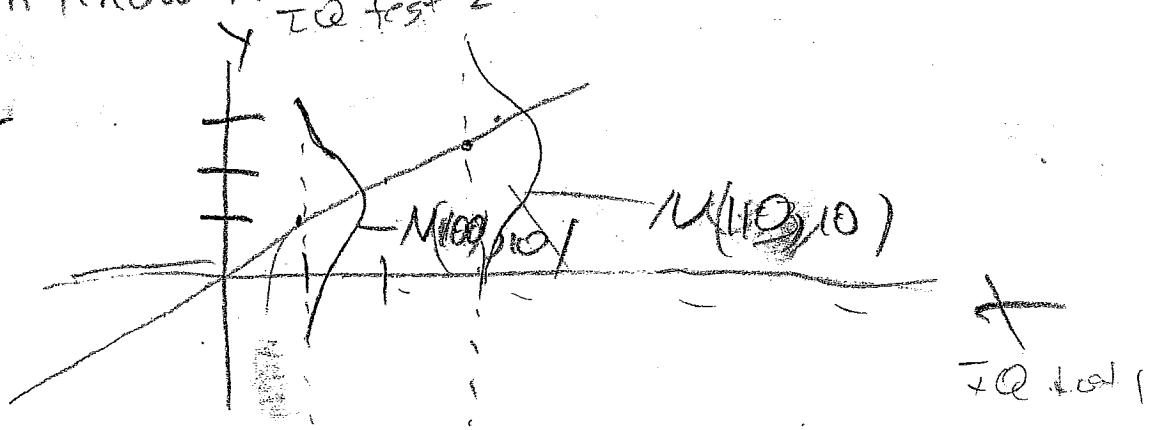
The population mean regression line is

$\mu_y = \alpha + \beta x$ = mean of y given that the value of the explanatory variable is x.

2) p532 $y_i = \alpha + \beta x + \varepsilon_i$

The $\varepsilon_i \overset{\text{independent}}{\sim} N(0, \sigma)$. α, β, σ are unknown.

3) p532



Interpretation: for every value x, the distribution of Y at x is $N(\mu_y, \sigma)$

$= N(\alpha + \beta x, \sigma)$.

Tunnel

4) p533 $s = \sqrt{\frac{1}{n-2} \sum (y-\hat{y})^2} = \sqrt{\frac{1}{n-2} \sum (\text{residual})^2}$

$= \sqrt{MSE}$

estimates σ

plug $\hat{y} = a + bx$ $\quad b = r \frac{s_y}{s_x}$ $\quad a = \bar{y} - b\bar{x}$

5) p533 ith residual $e_i = y_i - \hat{y}_i$

= observed response − predicted response

6) The normality assumption of the ε_i is important. The e_i estimate the ε_i.

7) *p537 100 c% CI for slope β

is $b \pm t^* SE_b$ where

$P(-t^* < t_{(n-2)} < t^*) = c$ $\quad df = n-2$

Get t^* from table C if df $\leq$

use $t^* = z^*$ from table C

or use output if df >

and SE_b would be given.

8) *p540 test H_0 $\beta = 0$

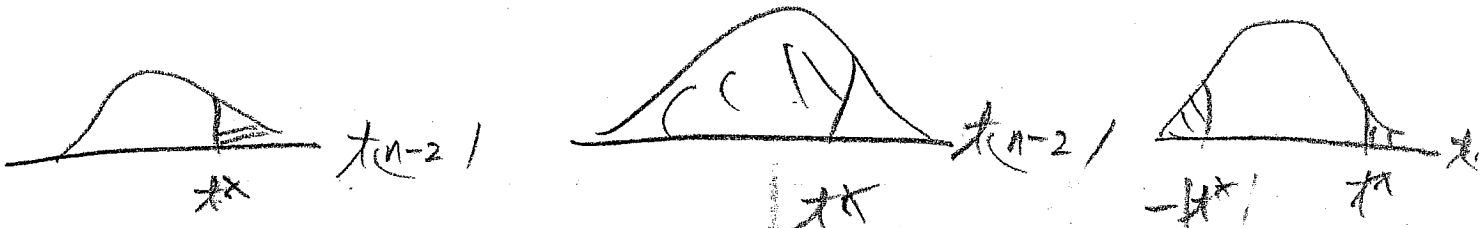
test statistic $t_0^* = \frac{b}{SE_b} \sim t_{(n-2)}$

$df = n-2$

P values

right tail | left tail | two tail

H_A $\beta > 0$



$P(t_{(n-2)} > t^*)$ $P(t_{(n-2)} \leq t^*)$ $2P(t_{(n-2)} \geq |t_0^*|)$

a) For inference we will always use output. Same 4 step test.

10) p544 100 c% CI for μ_y when $x = x^*$

is $\hat{y} \pm t^* SE_{\hat{\mu}_y}$

where $\hat{y} = a + bx^*$, $SE_{\hat{\mu}_y}$ is given in output

and $P(-t^* \leq t_{(n-2)} \leq t^*) = c$

11) p544 A prediction interval is for a new observation y_{new} rather than a mean. Analogy 68-95-99.7 rule.

$y_i - \hat{y}_i = e_i \approx \varepsilon_i \simeq N(0, \sigma)$

So there is about 95% chance

that y_{new} is in $(\mu_y - 2\sigma, \mu_y + 2\sigma)$ if y_{new} is from a $N(\mu_y, \sigma)$ pop.

12) p544 A $100(1-\alpha)\% = 100C\%$ prediction interval for a future observation y_{new} at an explanatory value of $x = x^*$ is

$$\hat{y} \pm t^* SE_{\hat{y}} \text{ where}$$

$$P(-t^* \leq t_{(n-2)} \leq t^*) = 1-\alpha = C.$$

Note y_{new} is a random variable, not a parameter.

13) p553 $SE_b = \dfrac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \to 0$ as $n \uparrow$

p544 $SE_{\hat{\mu}_y} = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}} \to 0$ as $n \uparrow$

p544 $SE_{\hat{y}} = s\sqrt{1 + \dfrac{1}{n} \dfrac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}} \to \sigma$ as $n \uparrow$

14) ex $n = 92$

	df	coef.	standard error	T	P (p. 171)
intercept	1	1.1630	0.0066	177.3	0.0000
log skin	1	$-0.0631 = b$	$0.0041 = SE_b$	$-15.2 = t^*$	0.0000 = p value

Want to predict $y =$ body density from

x = log skinfold $\equiv$ OM282 5

a 95% CI for β is

$b \pm t^* SE_b$ = $df = 92-2 = 90 > 30$

$-0.0631 \pm 1.96\,(.0041)$

$= -0.0631 \pm .0080 = (-0.071, -.055)$

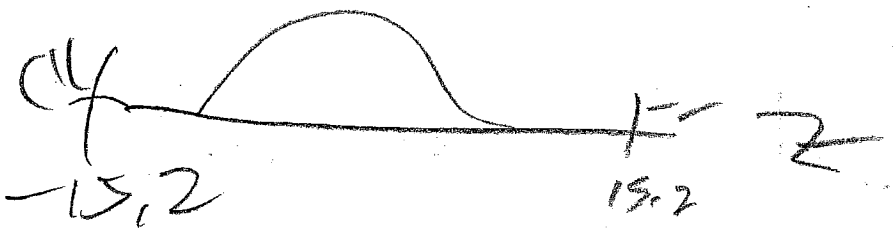
T-4 step test

1) H_0 $\beta = 0$ H_A $\beta \neq 0$

2) $T = -15.2$

3) Pval $= 2P(t(90) > 15.2)$

$df = 90 > 30$ use table A



$= 2(0) = 0.0$

or use output

4) Reject H_0 knowing the value of log skinfold helps predict body density.

ex) Same example but $n = 9$ §7.9

95% CI $df = n-2 \Rightarrow$ $t^* = 2.365$ table c

$b \pm t^* SE_b =$

$-0.0631 \pm 2.365(0.0041) =$

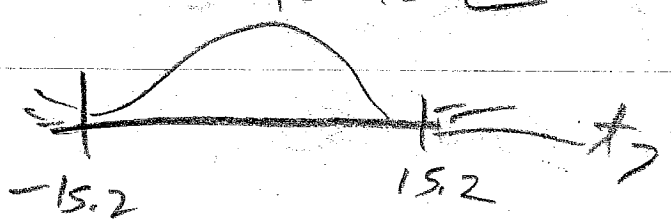
$-.0631 \pm .0097 = (-0.073, -0.053)$

4 Step test

s1) H_0 $\beta_1 = 0$ H_A $\beta_1 \neq 0$ $\alpha / .005$ $7 < 5.405$

s2) $t_0^* = -15.2$

s3) $df = 7$ pval =



≈ 0 (pval $< 2(.005) = .001$)

s4) reject H_0 knowing the value of log skinfold helps predict the body density.

ex FOM 282 5

$\hat{y} = 1.0892 + .1890x$

Find 95% CI for μ_y and 95% PI for y_{new} if $x^* = 20$, $n = 16$

$SE_{\mu_{\hat{y}}} = 0.0855$ $SE_{\hat{y}} = 0.3496$

Output

Fit	stdev fit	95% CI	95% PI
4.8692	0.0855	(4.6858, 5.0526)	(4.1193, 5.619

$\hat{y}$ $SE_{\mu_{\hat{y}}}$

$\hat{y} = 1.0892 + .1890(20) = 4.8692$

stdev fit $= SE_{\mu_{\hat{y}}}$

So 95% CI for $\mu_{\hat{y}}$ is

$\hat{y} \pm t^* SE_{\mu_{\hat{y}}}$ $df = n-2 = 14$

$t^* = 2.145$

$4.8692 \pm 2.145(0.0855) = 4.8692 \pm .018$

$= (4.6858, 5.0526) =$ 95% CI for $\mu_{\hat{y}}$

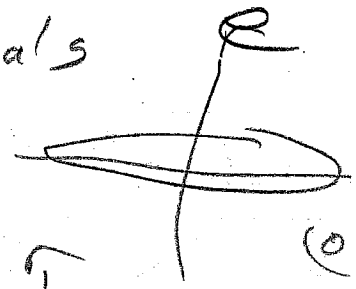
$SE_{\hat{y}} = 0.3496$ $4.8692 \pm 2.145(0.3496)$

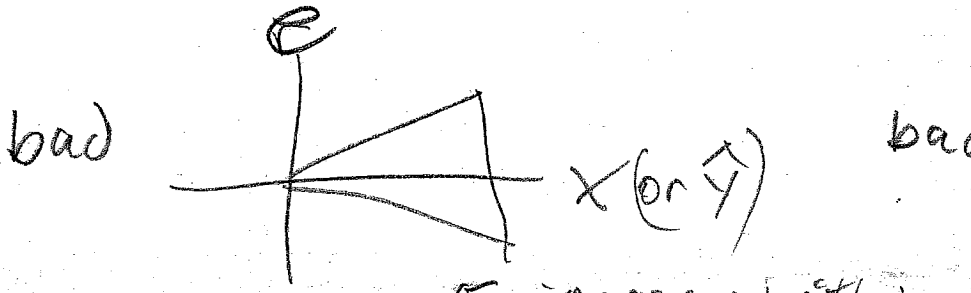
$= 4.8692 \pm .7499 = (4.1193, 5.6191) =$ 95% for

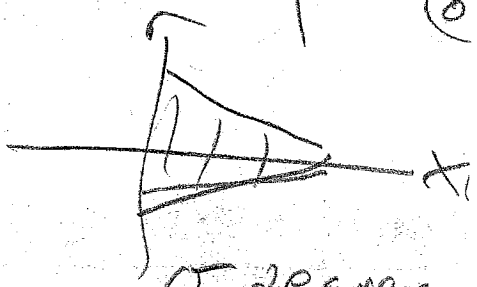
15 p546 i) scatter plot of Y vs x should be "football shape"
No curvature in a plot of e_i vs $\hat{Y}$ (residuals)

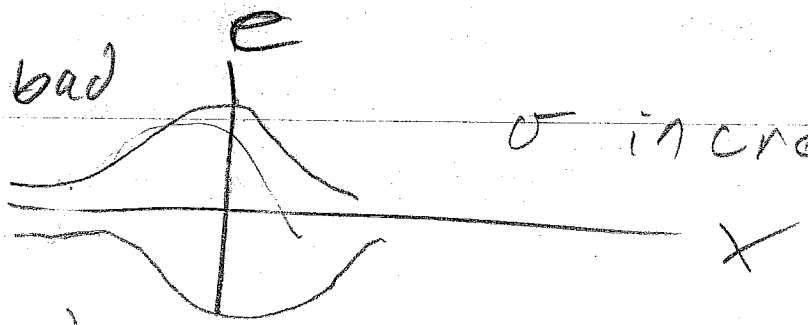
ii) $\varepsilon_i \sim N(0, \sigma)$ σ does not depend on x

The plot of the residuals vs x should look like





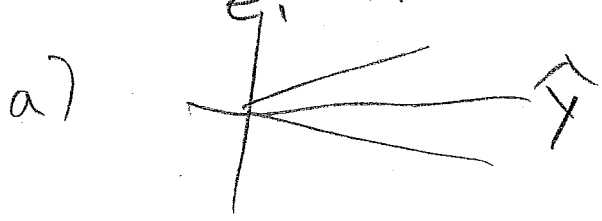




iii) $\varepsilon_i \sim N(0, \sigma)$ normality

make a histogram of the residuals e_i; check for outliers.

ex Fig 11.8 p549



$\Sigma e_i = 2$ [illegible]

§3.2 Experiments F 0282 59

1) p166 In an <u>experiment</u> the investigators decide which individuals get a "treatment" and which do not, then observe a response.

2) p166 In an <u>observational study</u>

(the individuals decide if they take the treatment) investigators simply observe a response.

3) Know the difference. Clue! the study is observational if the "treatment" eg smoking, illegal drugs, marriage divorce, etc is not something a person would do at a flip of a coin.

A study of a new medicine before FDA approval should always be an experiment. A study of "individuals" when the "individuals" are not people and the treatment is cheap, the study will often be an experiment.

A study of 2 established medicines could be observational or an experiment.

ex Effect of smoking on cancer obs

ex effect of Viagra on impotence when the drug was being made exp

ex study of people with 3 or more marriages ob. interracial marriage ob. heavy alcohol consumption ob.

ex effect of irrigation fertilizer and herbicide on crop yields exp

ex effects of smoking on rats exp

§9.3

4) p186 The "individuals" on which an experiment is done are called experimental units. They could be people, rats, ~~farms~~ fields etc. Each unit receives a treatment.

5) p187 The purpose of an experiment is to reveal the response of one variable as explanatory variables change.

ex sick people give medicine or sugar pill
response was person cured or not
explanatory medicine vs sugar pill

6) we compare the treatment with other treatments or a placebo (sham treatment p189)

7) p189, 193 The most important step in an experiment is randomization, ~~of subjects into 2 or more~~ groups, each group gets a different trt or placebo. We know use random numbers (impersonal chance) to decide which units get the treatment and which units get the placebo. see point 17

Analogy (probability samples) use random numbers F0282 6D
to select a sample. Probability samples are good voluntary response samples are bad.

8) (p189) The control group is the group that receives the placebo or standard treatment. Compare the treatment group with the control group.

9) Using randomization to decide who gets in the treatment group and who gets in the control group minimizes the effects of lurking variables.

(10) i) best — double blinded completely randomized controlled expt ← used: medicines, vaccines
ii) next best — single " " " "
iii) good — completely randomized controlled expt ← surgery, AG, industry
iv) OK — observational ← ethically can't divide people into 2 groups (smoking, alcohol)
Know v) terrible — controlled, not randomized
vi) terrible — just a trt group, no control group

p189 Double blinded means neither the patients nor the doctors know who got the treatment and who got the placebo.
Completely randomized controlled means there is a treatment group and a control group and random numbers were used to decide who gets in which group.

ex) Salk vaccine for polio

2 studies NFIP (controlled not randomized) and a double blinded randomized controlled experiment (saline solution as placebo),

Think of factors that might make contracting polio more likely

treatment group control group

income
race
gender
hygeine
hereditary
# of siblings

For all the lurking variables you can think of, there are many more that you did not think of. Randomization causes the treatment and control groups to have approximately the same proportion of any lurking variable. Eg if 15% of the units are white female with brown hair, then about 15% of the treatment units and about 15% of the control units are too. However chance variation means you could get 14.7% or 15.3% say, (As more units are used) the chance variation decreases.

The salk vaccine could only be given to children if there parents gave consent

DB CRC Expt

		size	rate per 100000
expt	trt	200000	28
	control	200000	71
	no consent	350000	46

NFIP

	size	rate
grade 2 tot	225000	25
grade 1 & 3 control	725000	54
grade 2 & no consent	125000	44

not part of expt, but recorded anyway

The NFIP study was badly done,
NO randomization
Treatment group was all grade 2 consenters.
The control group was all grade 1 & 3 students and contained many kids whose parents would not have given consent. Consenters are different than nonconsenters, Polio is one of the few diseases that affects the rich more than the poor, but a greater percentage of poor did not give consent, Children in poorer families tend to get earlier milder cases of polio.

11) P203 A study is biased if it systematically favors certain outcomes. In the NFIP study the control group was less

likely to get polio than the treatment group because the treatment group only had 2nd grade consenters while the control group had all 1st and 3rd graders (waddle consenters and nonconsenters), NFIP was biased against vaccine (615)

(2) p189 placebo effect — Idea of treatment gives relief

25/54 doctors not look as good as 29/21

ex gel capsule or medicine for headache relief

If one group is given medicine and the other nothing, can't tell how much relief is due to idea of treatment and how much to the medicine (confounding)

To defeat confounding give one group medicine, one group placebo then compare.

ex cirrhosis of liver portacaval shunt

51 studies 4 randomized controlled(1)

expt	degree of enthusiasm: marked	moderate	none
i) no controls	24	7	1
ii) controls not randomized	10	3	2
iii) randomized controlled	0	1	3

Some patients were too sick for surgery. The randomized controlled experiment divided patients into eligible and uneligible patients.

Half the eligible patients got the OE 2.5.2 treatment half did not.

In ii) doctors decided who got treatment and tended not to give the treatment to the sicker patients. There study is biased in favor of the treatment

13)

good experiments	bad
double blinded randomized controlled single blinded randomized controlled randomized controlled	not controlled not randomized controlled.

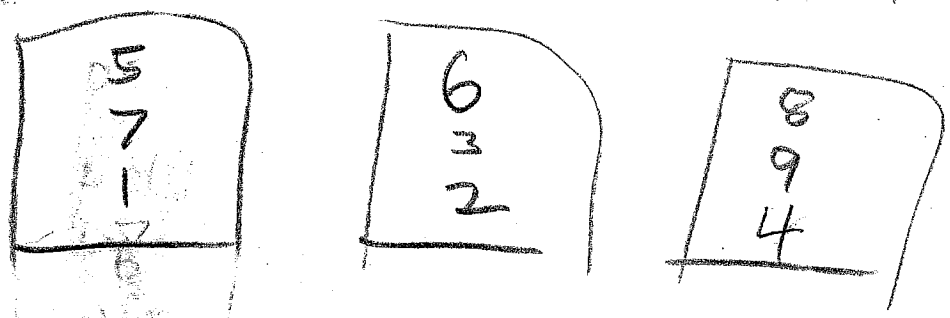
14) Often can't use placebo or blinding ex surgery doctor and patient know if patient got surgery) can't pull out an appendix as a placebo for liver surgery.

15) P206 There can be more than 2 treatments. When all experimental units are allocated at random among all treatments, the design is called completely randomized

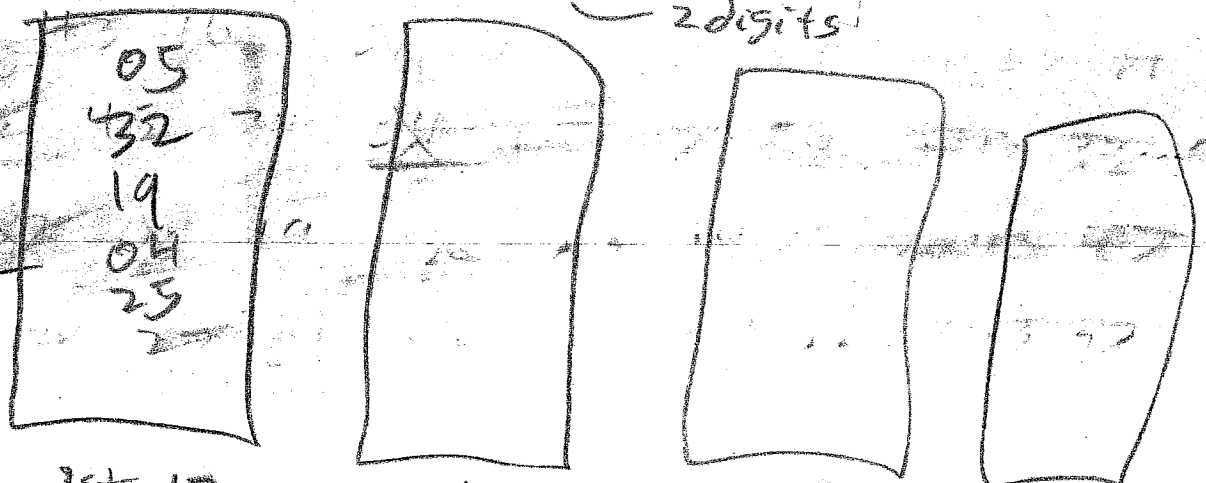
16) ex 3 treatments 19 units
1 digit

625

line 131 05 00 71 66 32 81 19 41 48 73
done
04 17 78 55 76 45 19 59 65 65
[illegible] 10 73 25 52 59 84 29 20 87 96



ex
4 treatment 40 units
2 digits



1st 10 next 10 next 10 last 10

line 131 05 00 71 66 32 81 19 41
48 73 04 19 78 55 76 45 19 59 65 65

line 132 68 73 25 52 etc

17) Randomization minimizes the effects of lurking variables makes the groups as alike as possible except for different treatments, helps prevent biases of the experimenter from influencing the outcome of the exp, and causes $\bar{x}_0$ or $\hat{p}_0$ to be approximately normal