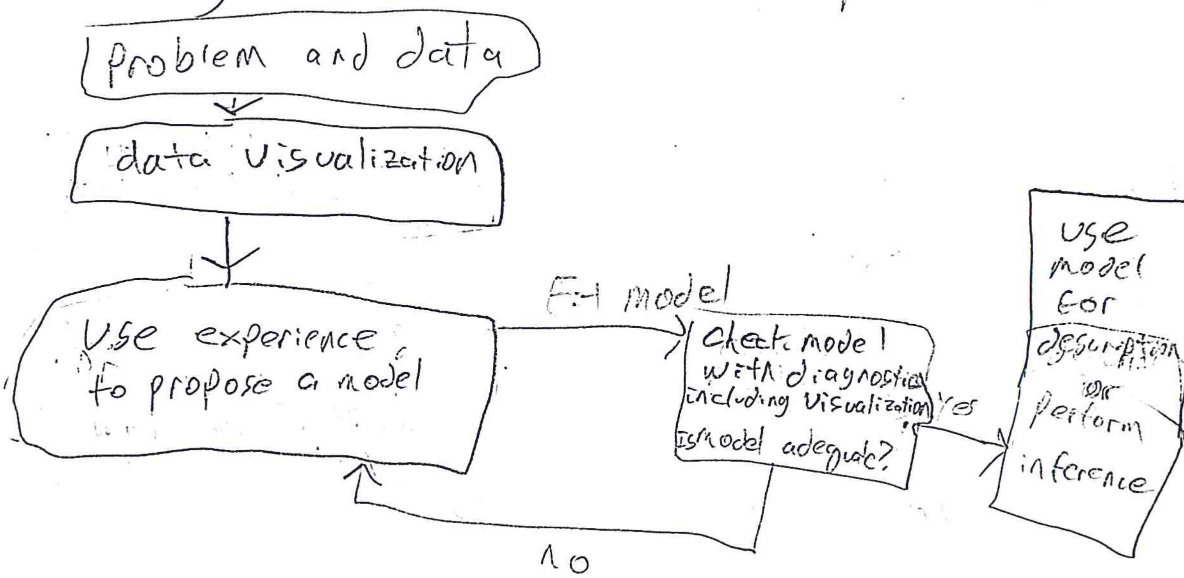
1] **Statistics** is the science of obtaining useful information from data.

2] P.1 A **statistical model** is used to provide a useful approximation to the population that generated the data.

(A parametric location model)

ex] $Y_i = \mu + e_i \quad i = 1, \ldots, n$ where the e_i are iid $N(0, \sigma^2)$. So the Y_i are iid $N(\mu, \sigma^2)$. Confidence intervals for μ and hypothesis tests for $H_0\ \mu = \mu_0$ vs $H_A\ \mu \overset{>}{\underset{<}{\neq}} \mu_0$ should be familiar from intro stat courses.

3] Model building is an iterative process.



4] P4 **Robust statistics** can give useful results when the model holds and when a certain specified model assumption is incorrect.

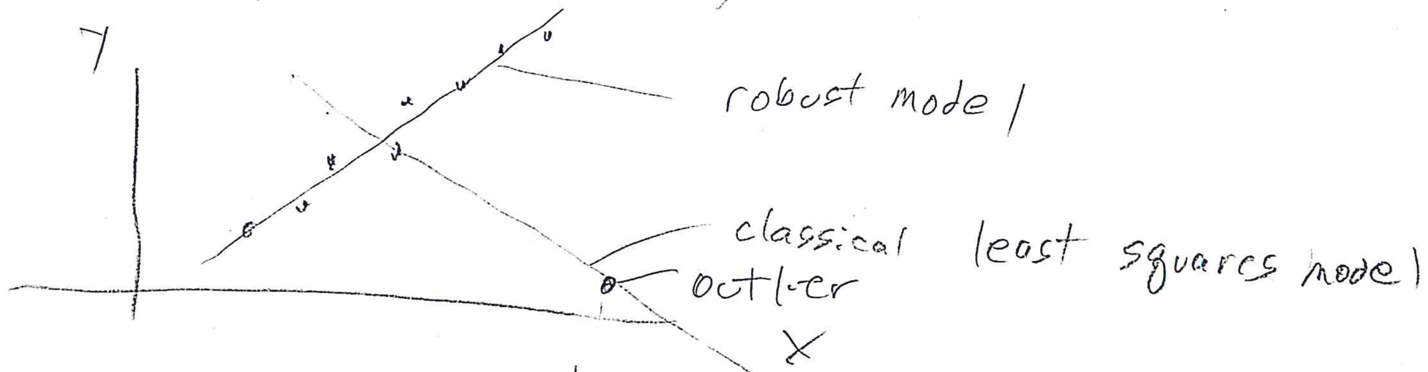
5) One assumption violation is the

presence of outliers: observations

far from the bulk of the data.

(often due to recording errors, not always bad eg spouse, good teachers doctors etc)

ex $Y =$ height $\quad X =$ height at shoulder



we are also interested in methods that are robust to the assumption of a parametric dist, eg $\bar{Y} \approx N(\mu, \frac{\sigma^2}{n})$ if $Y_1, \ldots, Y_n$ are iid with $E(Y) = \mu$ $V(Y_i) = \sigma^2$.

6) p2 In a 1D regression, the response variable Y that you want to predict with a vector $\underline{x} = (x_1, \ldots, x_p)^T$ of predictor variables is conditionally independent of $\underline{x}$ given $h(\underline{x})$ written $Y \perp\!\!\!\perp \underline{x} \mid h(\underline{x})$, where $h(\underline{x})$ is the sufficient predictor and $\hat{h}(\underline{x})$ is the estimated sufficient predictor. A response plot is a plot of ESP vs Y. ESP

7) The single index model

or $Y_i = m(\underline{x}_i^T \underline{\beta}) + e_i$ is a 1D model, where e_i is an error, eg $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$.

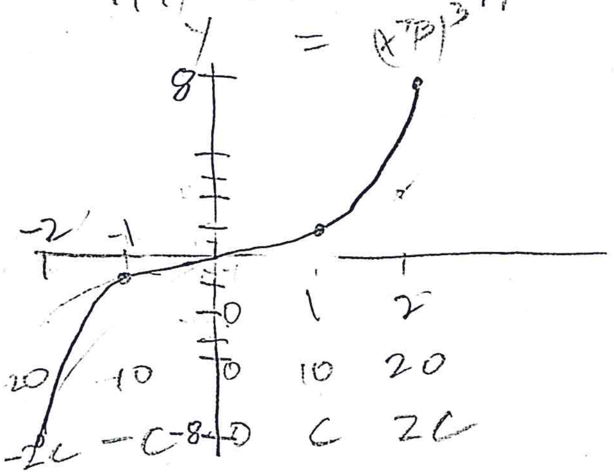
8) Another assumption violation is that m is unknown or misspecified, see ch. 9

c) Suppose $Y = m(\underline{x}^T\underline{\beta})$ horiz axis vertical axis

i) what happens if you plot $\underline{x}^T\underline{\beta}$ vs Y ?

ii) what happens if you plot $c\underline{x}^T\underline{\beta}$ vs Y for $c>0$?

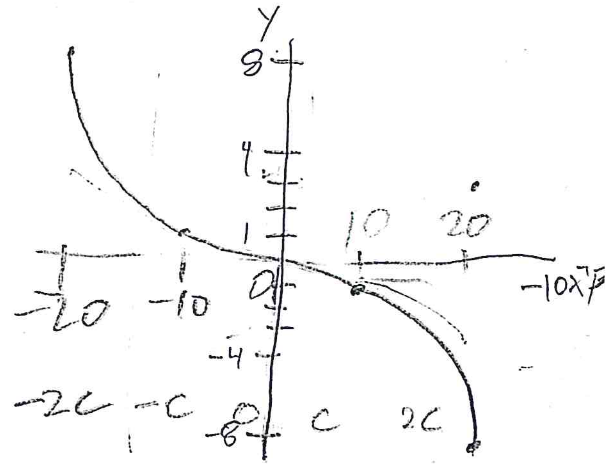
iii) what happens if you plot $c\underline{x}^T\underline{\beta}$ vs Y for $c<0$?



$x^T\beta$	y	$10x^T\beta$
-2	-8	-20
-1	-1	-10
0	0	0
1	1	10
2	8	20

$x^T\beta = w$

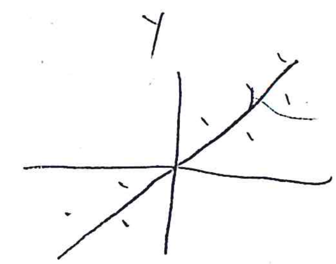
$10\,x^T\beta$



If $Y = m(\underline{x}^T\underline{\beta}) + e$ and $\underline{\beta}$ was known a plot of $c\underline{x}^T\underline{\beta}$ vs Y lets you "see m" up to error.

Ex} $Y = \underline{x}^T\underline{\beta} + e$ let $w = x^T\beta$

$Y = \underbrace{0 + 1w}_{m \text{ is the line through the origin with unit slope}} + e$



$Y_i - x_i^T\beta = e_i$ is the vertical deviation

$x^T\beta$

Note that $m(x^T\beta) = m\left(\frac{a + cx^T\beta - a}{c}\right) = m_{a,c}(a + cx^T\beta)$

where $m_{a,c}(u) = m\left(\frac{u-a}{c}\right)$.

So a plot of $a + cx^T\beta$ vs Y "shows m."

2.5

9] Idea: if $\hat{\beta}$ is a good estimator of $c\beta$ for some $c \neq 0$, then a plot of $\underline{x}^T\hat{\underline{\beta}}$ vs Y is almost a plot of $cx^T\beta$ vs Y.

10] Often the least squares estimator

$\hat{\underline{\beta}} = c\underline{\beta} + \underline{u}$ where the bias vector

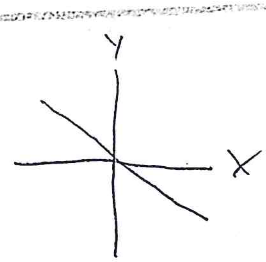
$\underline{u} = 0$ or is small, Often $\underline{u}$ can be made small by computing least squares on a subset of the data.

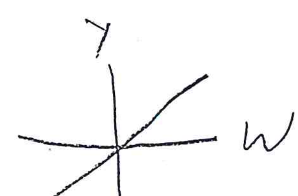
11] There are "useful models" but no "true model."

12] With a single predictor, the model $Y = m(x) + e$ can be visualized with a scatterplot of x vs Y. For a 1D model $Y = m(x^T\beta) + e$, m can be visualized with a scatterplot of $x^T\hat{\beta}$ vs Y if $\hat{\beta}$ is a good estimator of $c\beta$ for $c \neq 0$.

ex) $Y = -X$ so $m(x) = -x$

Y	1	0	−1	2
$W = -X$	1	0	−1	2
X	−1	0	1	2





for predicting Y, W is about as good as X although the "true m" is flipped about the Y axis.

1) p19* The location model is $Y_i = \mu + e_i$, $i = 1, \dots, n$.

2) p19 know An important robust technique for the location model is to make a plot of the data.

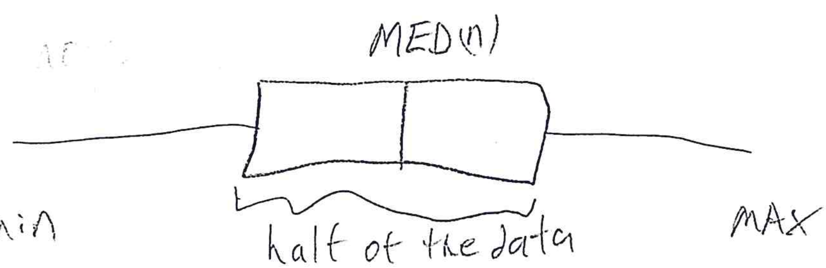
3) common assumption: $e_1, \dots, e_n$ are iid from a distribution with 0 mean and variance σ^2 and n is large enough so that the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \approx N(\mu, \frac{\sigma^2}{n})$, ie the central limit theorem CLT holds.

4) A <u>dot plot</u> is a plot of i vs Y_i

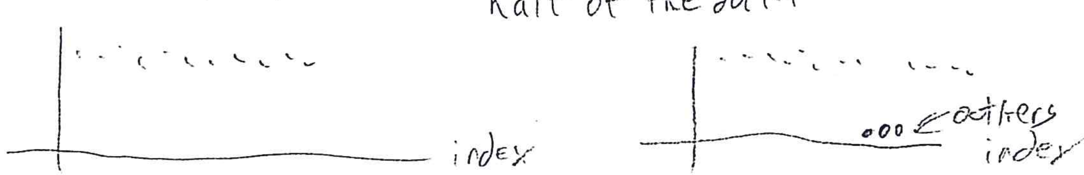
A <u>histogram</u> tries to approximate the probability density function (pdf) $f(y)$ of a continuous random variable (RV) Y and to approx the probability mass function (pmf) $P(Y=y)$ of a discrete RV Y.

A <u>density estimate</u> is a smoother approx for $f(y)$.

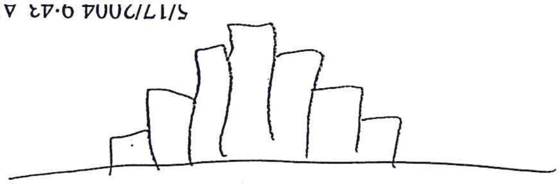
A typical <u>boxplot</u> summarized the dot plot.



ex) Y = height
dot plot

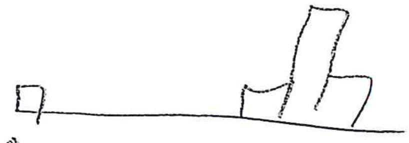


3.5

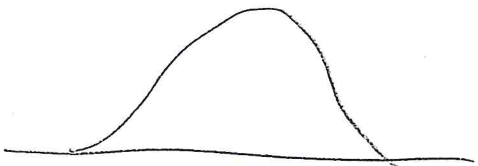


histogram normal data

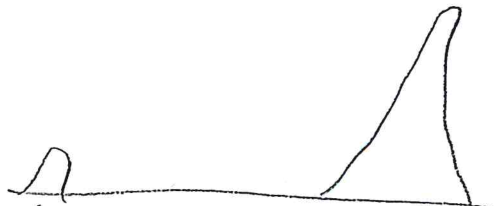
+ outliers



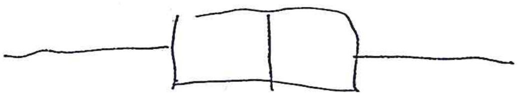
outliers



density estimate for normal data



outliers



box plot

outliers

5) Know p20-22 the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

If the data is $Y_1, \ldots, Y_n$, then the order statistics are $Y_{(1)}, \ldots, Y_{(n)}$ where $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$ are the Y_i's written in ascending order.

min max

The sample variance $S^2 = Var(n) = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \dfrac{\sum_{i=1}^{n} Y_i^2 - n(\bar{Y})^2}{n-1}$

The sample standard deviation $S = \sqrt{Var(n)}$

The sample median $MED(n) = \begin{cases} Y_{(\frac{n+1}{2})}, & n \text{ odd} \\ \dfrac{Y_{(\frac{n}{2})} + Y_{(\frac{n}{2}+1)}}{2}, & n \text{ even} \end{cases}$

The sample median absolute deviation $MAD(n) = MED|Y_i - MED(n)|$

That is, let $D_i = |Y_i - MED(n)|$, then $MAD(n)$ is the median of $D_1, \ldots, D_n$.

6) Know for Final: given data compute

$\overline{Y}$, S, MED(n) and MAD(n)

ex) Also see HW1 problem 2.10

Consider the data set 66, 3, 8, 5, 2.

Find a) $\overline{Y}$ b) S c) MED(n) d) MAD(n)

Soln) a) $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{84}{5} = \boxed{16.8} = \frac{66+3+8+5+2}{5}$

b) $S^2 = \frac{\sum Y_i^2 - n(\overline{Y})^2}{n-1} = \frac{4458 - 5(16.8)^2}{4} = \frac{3046.8}{4}$

$S^2 = 761.7$

$S = \sqrt{S^2} = \sqrt{761.7} = \boxed{27.5989 = S}$

(Don't forget to square $\overline{Y} = 16.8$.)

c) Sort data 2, 3, 5, 8, 66

$\boxed{MED(n) = 5}$

d) $Y_i - MED(n)$: −3, −2, 0, 3, 61

Sort $|Y_i - MED(n)|$ 0, 2, 3, 3, 61

$\boxed{MAD(n) = 3}$

$= \sum Y_i - \sum Y_i = 0$ 4.5

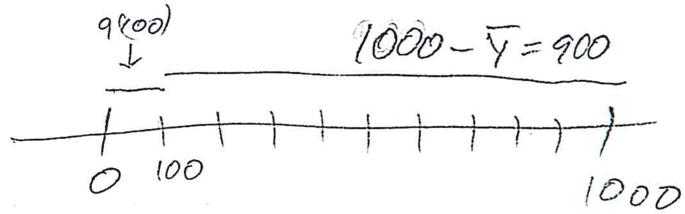
So $\bar{Y}$ is the value such that the sum

sum of the distances of the Y_i's $< \bar{Y}$ " = " $>$

So outliers affect $\bar{Y}$

(Y_i's $= \bar{Y}$ contribute 0 to the sum)

ex $\underbrace{0, 0, 0, 0, 0, 0, 0, 0, 0,}_{9\ 0\text{'s}}$ 1000



$\bar{Y} = 100$ is not a typical value

$$\overbrace{1000 - \bar{Y}}^{\text{deviations from } Y > \bar{Y}} = \overbrace{9(\bar{Y} - 0)}^{Y\text{'s} < \bar{Y}} = 900$$

8) MED(n) is such that

at least half of the Y_i's $\le$ MED(n) and

at least half of the Y_i's $\ge$ MED(n)

ex) In the last ex, MED(n) = 0, a typical value

1, 2, 3 — MED(n) = 2

1, 2, | 3, 4 — MED(n) = 2.5

1, 2, 3, 4, 5

can replace these by any numbers greater than 3, and MED(n) does not change

replace this by any number x

	MED(n)
$-\infty < x \le 3$	3
$3 < x \le 4$	x
$4 < x < \infty$	4

MED(n) is the "most" outlier resistant estimator of location (MAD mad)

9) $S^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2 \approx$ sample mean of $(Y_i - \bar{Y})^2$

not robust! a single outlier greatly changes S^2 and S

but $MAD(n) = \text{median}\, |Y_i - MED(n)|$ is the "most" outlier resistant measure of spread.

10) pop quantities $E(Y)$ $Var(Y)$ $MED(Y)$ $MAD(Y)$

$E(Y) = \mu$, $Var(Y) = \sigma^2$

sample analog $\bar{Y}$ S^2 $MED(n)$ $MAD(n)$

11) *p 23 The population median $MED(Y)$ is any value such that $P(Y \le MED(Y)) \ge 0.5$ and $P(Y \ge MED(Y)) \ge 0.5$

12) * pop $MAD(Y) = MED(|Y - MED(Y)|)$

13) p 23 Let $f_Y(y)$ be the pdf of Y.

a) The family of pdf's $f_W(w) = f_Y(w - \mu)$ is the location family for $W = \mu + Y$, $\mu \in \mathbb{R}$.

b) The family of pdfs $f_W(w) = \frac{1}{\sigma} f_Y\left(\frac{w}{\sigma}\right)$ is the scale family for $W = \sigma Y$, $\sigma > 0$.

c) The family of pdfs $f_W(w) = \frac{1}{\sigma} f_Y\left(\frac{w-\mu}{\sigma}\right)$ is the location scale family for $W = \mu + \sigma Y$ where $\mu \in \mathbb{R}$ and $\sigma > 0$

Note: $F_W(w) = P(W \le w) = P(\mu + \sigma Y \le w)$

$$= P\left(Y \le \frac{w-\mu}{\sigma}\right) = F_Y\left(\frac{w-\mu}{\sigma}\right) = T$$

So $f_W(w) = \frac{d}{dw} F_Y\left(\frac{w-\mu}{\sigma}\right) = \frac{1}{\sigma} f_Y\left(\frac{w-\mu}{\sigma}\right)$.

14) p333 <u>know for E1</u> The cdf $F_Y(y) = P(Y \le y)$.

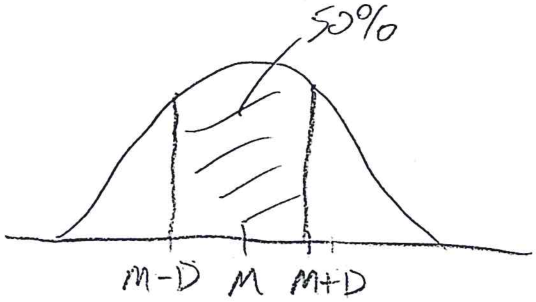
a) Let $M = MED(Y)$. To find M, solve $F_Y(M) = 0.5$.

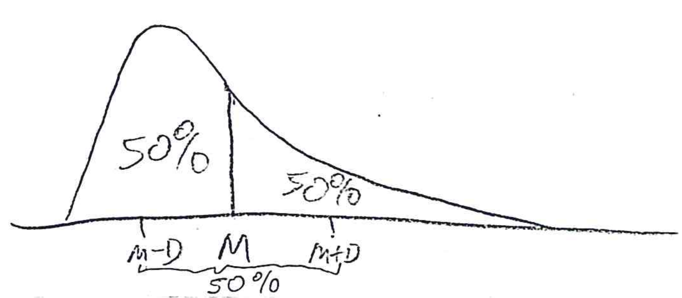
b) Let $D = MAD(Y)$. After finding M, find D by solving $F_Y(M+D) - F_Y(M-D) = 0.5$, often numerically.

c) If $W = \mu + \sigma Y$, then $MED(W) = \mu + \sigma M$ and $MAD(W) = \sigma D$.

d) If Y has a pdf that is symmetric about μ, then $MED(Y) = \mu$ and $MAD(Y) = y_{0.75} - MED(Y)$

where $P(Y \le y_\alpha) = \alpha$, ie $y_{0.75}$ is the 75th percentile of Y.





... Suppose Y is a RV with a symmetric pdf f_Y and cdf $F_Y(y) = \begin{cases} 0 & y \le \theta_1 \\ \dfrac{y-\theta_1}{\theta_2-\theta_1} & \text{for } \theta_1 \le y \le \theta_2 \\ 1 & y \ge \theta_2 \end{cases}$

Find a) MED(Y) b) MAD(Y)

Soln a) $F_Y(M) = \dfrac{M-\theta_1}{\theta_2-\theta_1} \overset{set}{=} 0.5$

or $M = \dfrac{\theta_2-\theta_1}{2} + \theta_1 = \dfrac{\theta_2-\theta_1+2\theta_1}{2} = \dfrac{\theta_1+\theta_2}{2}$

b) Let $U = y_{.75}$

So $F_Y(U) = \dfrac{U-\theta_1}{\theta_2-\theta_1} = 0.75$

or $U = (\theta_2-\theta_1)\frac{3}{4} + \theta_1$

and $MAD(Y) = U - M = \frac{3}{4}(\theta_2-\theta_1) + \dfrac{4\theta_1}{4} + \dfrac{-2\theta_1-2\theta_2}{4}$

$= \dfrac{3\theta_2-3\theta_1+4\theta_1-2\theta_1-2\theta_2}{4} = \dfrac{\theta_2-\theta_1}{4}$

15) If Y is from a 2 parameter family [illegible]

$\mu = c_1 E(Y)$ and $\tau = c_2 VAR(Y)$,

then the method of moments estimator

is $\left(\hat{\mu} = c_1 \bar{Y}, \ \hat{\tau} = c_2 \frac{n-1}{n} S^2\right)$.

16) Know for E1: P27 **The MAD Method!**

if Y is a 2 parameter family

with $\theta = g_1(MED(Y), MAD(Y))$ and

$\lambda = g_2(MED(Y), MAD(Y))$ then

$\hat{\theta} = g_1(MED(n), MAD(n))$

$\hat{\lambda} = g_2(MED(n), MAD(n))$.

ex) $Y \sim N(\mu, \sigma^2)$

$MED(Y) = \mu$ $\quad MAD(Y) \approx 0.6745\sigma$

so $\hat{\mu} = MED(n)$ and $\hat{\sigma} \approx \frac{MAD(n)}{0.6745} \approx 1.483 MAD(n)$.

ex) $Y \sim C(\mu, \sigma)$ cauchy $E(Y)$ and $Var(Y)$ do not exist, but $MED(Y) = \mu$ and $MAD(Y) = \sigma$

so $\hat{\mu} = MED(n)$ and $\hat{\sigma} = MAD(n)$.

17) Consider "intervals" that contain c cases $[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], \ldots, [Y_{(n-c+1)}, Y_{(n)}]$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, \ldots, Y_{(n)} - Y_{(n-c+1)}$. Then $\text{shorth}(c) = [Y_{(s)}, Y_{(s+c-1)}]$ is the closed interval with the shortest length.

know for E1

ex) Let $c = 4$. Data below has $n = 7$.

0, 1, 3, 6, 9, 10, 11

intervals containing $c = 4$ cases:

$6 = 6 - 0$

$8 = 9 - 1$

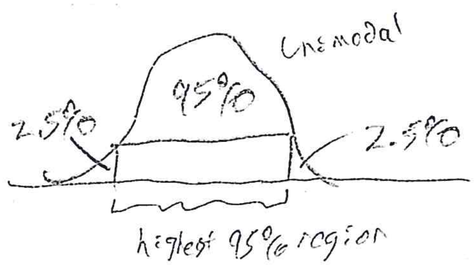
$7 = 10 - 3$

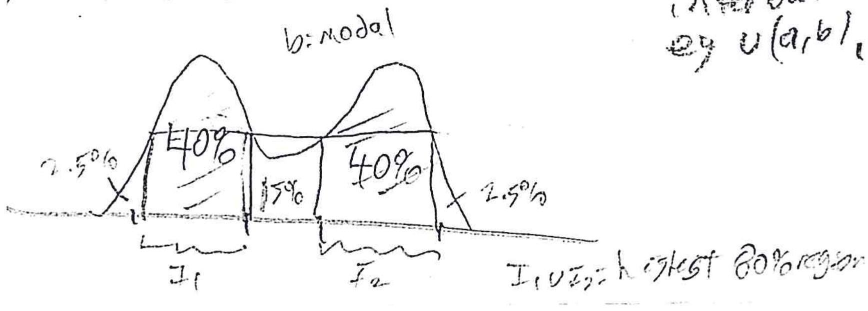
$5 = 11 - 6$

$\boxed{[6, 11] = \text{shorth}(4)}$

5 is shortest length

18) The highest $100(1-\delta)\%$ density region of a pdf is found by moving a horizontal line down from the top of the pdf so that the line intersects the pdf at one or more intervals and the sum of the areas under the pdf corresponding to the intervals $= (1-\delta)$. The pdf can't have a positive flat interval eg $U(a,b)$.





The shorth(c)

estimates the highest density $100(1-\delta)\%$ region if that region is an interval. Then the shorth(c) estimator can be used as a $100(1-\delta)\%$ large sample prediction interval (PI).

If $Y_1, ..., Y_n, Y_f$ are iid where Y_f is a future observation, $[L_n, U_n]$ is a $100(1-\delta)\%$ large sample PI if $P[Y_f \in [L_n, U_n]]$ is eventually bounded below by $1-\delta$ as $n \to \infty$. often want $P(Y_f \in [L_n, U_n]) \to 1-\delta$ as $n \to \infty$

The most used statistical PI's assume $Y_i \sim N(\mu, \sigma^2)$ and have coverage much smaller than $100(1-\delta)\%$ eg 95% or 50%. Let $k_1 = \lceil \frac{n\delta}{2} \rceil$, $k_2 = \lceil n(1-\frac{\delta}{2}) \rceil$. $[Y_{(k_1)}, Y_{(k_2)}]$ is the nonparametric PI.

24] ch 9, ch 10 p 47 Concentration: Start with estimator T_0.

move Find the "half set" of cases closest to T_0

Y_i such that $|Y_i - T_0| \le MED|Y_i - T_0|$. Let (T_1, S_1^2) the sample mean and variance of these cases. (T_1, S_1^2).

Iterate to obtain $(T_1, S_1^2), ..., (T_k, S_k^2)$, Could iterate until convergence. If k is fixed, eg $k=10$, don't need to compute S_i^2.

Fact: $S_i^2 \le S_{i-1}^2$ and convergence occurs when $S_i^2 = S_{i-1}^2$.

The MB estimator uses $T_0 = MED(n)$.

The DGK estimator uses $T_0 = \bar{Y}$

When iterated to convergence, the half set seems to be estimating the same thing as the shorth $(c \approx \frac{n}{2})$ estimator,