Robust estimation § 5.5 , 6.5

- single sample , want 'improved' estimates of location , scale (instead of $\bar{y}$, s)

- 'improved' : not sensitive to outliers
  doesn't change dramatically
  if an outlier is added
  to data (or already there)

- example $\underset{i=1,\dots,n}{\text{median}}(y_i)$ to estimate ~~[illegible]~~ the location

  instead of $\bar{y}$

- example $\text{mad}(y_i) = \text{median}|y_i - \text{med}(y)| \times 1.4826$ ~~1.3445~~ built in

  instead of $s = \sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2}$

  OR $IQR = y_{(.75)} - y_{(.25)}$

  interquartile range

  to estimate scale use $IQR/1.394$

  not

Math: $y_i \sim \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$

$$y_i = \mu + \sigma e_i \qquad e_i \sim f(e)$$

Model for data, called location & Scale model

ex 1) $f(e) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}e^2}$ normal model

2) $f(e) = c_v \frac{1}{(1+e^2/v)^{\frac{v+1}{2}}}$? t_v-dist
t density with v degrees of freedom

⋮

lots

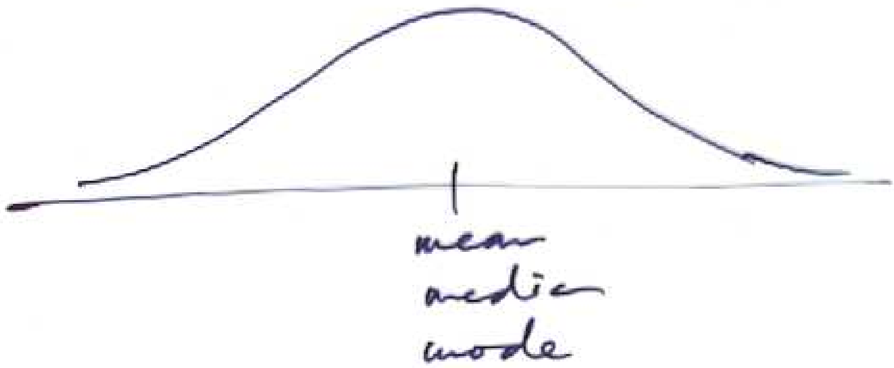
In 1) $\mu = EY \quad \sigma^2 = var(Y)$

but in 2) $\mu = EY$ but $\sigma^2 \neq var(Y)$
σ still measures 'spread'

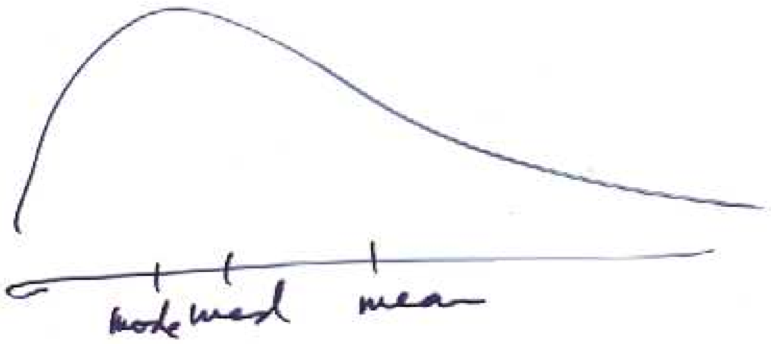
The factors 1.48 & 1.34 in MAD & IQRs ensure that $\binom{MAD}{IQR}$ estimates σ if $Y \sim$ Normal

$[MAD \xrightarrow{P} \sigma \quad n \to \infty]$

Question: what are we estimating if $f(e)$ is not symmetric?



symmetric



which one is μ?

- you should use robust estimators when $f(\cdot)$ is symmetric

- e.g. possibility of some large (or small) outliers, but in both directions

(not for chem, so ... ?)

median, IQR, MAD are examples of robust estimators (not affected by outliers)

a drawback - not very efficient (throws away data)

variance of median (in the limiting dist^n) is large, compared to variance of $\bar{y}$, if data is normal

a. var (median) = $\frac{1}{4f^2(0)}\frac{\sigma^2}{n}$ /

at normal $f(e) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}e^2}$

$f(0) = \frac{1}{\sqrt{2\pi}}$ $f^2(0) = \frac{1}{2\pi}$

& var (median) = $\frac{\sigma^2}{n}\left(\frac{2\pi}{4}\right)$ ← where's n? ntbc

var (mean) = σ^2/n

If we have 2 estimators of μ,

$$\text{efficiency} = \frac{\text{var (1st estimator)}}{\text{var (2nd estimator)}}$$

usually make do with variance of limiting distribution of the estimator

- point of robust est$^{\underline{n}}$ is to find estimate with good efficiency, but resistant to outliers

$Y_1, \ldots, Y_n$ iid $\frac{1}{\sigma} f\left(\frac{y_i - \mu}{\sigma}\right)$

likelihood f$^{\underline{n}}$ $L(\mu, \sigma; y) = \prod_{i=1}^{n} \frac{1}{\sigma} f\left(\frac{y_i - \mu}{\sigma}\right)$

log-lik f$^{\underline{n}}$ $\ell(\mu, \sigma) = \sum \log f\left(\frac{y_i - \mu}{\sigma}\right) - n \log \sigma$

max. lik. est. for μ (σ known)

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{solve for } \mu$$

$$(1) \quad \sum_{i=1}^{n} g\left(\frac{y_i - \mu}{\sigma}\right) = 0 \quad \text{solve for } \mu$$

$$g(\cdot) = (-\log f)'$$

if $f(e) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}e^2}$

$$\log f = -\tfrac{1}{2}e^2 - \text{const}$$

$$(-\log f)' = e$$

(1) $\Rightarrow \sum \left(\frac{y_i - \mu}{\sigma}\right) = 0 \Rightarrow \hat{\mu} = \bar{y}$

~~"Robust choices of g" include~~

Instead of (1), we'll <u>define</u> $\tilde{\mu}$ by

(2) $\sum_{i=1}^{n} \psi\left(\frac{y_i - \mu}{\sigma}\right) = 0$ for some choice $\psi(\cdot)$

p. 122, 123

$$\psi(x) = \begin{cases} -c & x < -c \\ x & -c \le x \le c \\ c & x > c \end{cases}$$

tuning constant c

<u>M-estimate</u> Huber's M-estimate $\longleftrightarrow$ huber(chem)

estimator is mean of middle with "outliers" replaced by c

$\psi = \rho'$ $\rho = \begin{cases} x^2 & |x| \le c \\ \text{linear} & |x| > c \end{cases}$

$c \overset{1-1}{\Leftrightarrow}$ scale σ (known)

$$\psi(x) = \begin{cases} -c \\ x \\ c \end{cases} \qquad \sum \psi\left(\frac{y_i - \mu}{\sigma}\right) = 0$$

$$\Longleftrightarrow \left.\begin{array}{l} \min\limits_{\mu} \sum \rho\left(\frac{y_i - \mu}{\sigma}\right) \text{ where} \\ \\ \rho'(\,) = \psi(\,) \end{array}\right\}$$

i.e. $$\rho = \begin{cases} x^2 & |x| < c \\ c(\underline{2|x| - c}) & |x| > c \end{cases}$$

$$\min_a \sum_{i=1}^{n} |x_i - a| \qquad \min \sum (x_i - a)^2$$