4/9/09 Math 152.

Median

defined as the middle value of the ordered observations

if the sample size is even, the median is the average of the two middle values

The median is a "robust" measure of location since changes in the extreme values do not effect its value

If the data are a sample from a continuous distribution the sample median is an estimate of the population median η.

In this case, we have

Confidence intervals for the pop. median.

Such an interval will have the form $(X_{(k)}, X_{(n-k+1)})$

$$P(X_{(k)} \leq \eta \leq X_{(n-k+1)})$$

$$= 1 - P(\eta < X_{(k)} \text{ or } \eta > X_{(n-k+1)})$$

$$= 1 - P(\eta < X_{(k)}) - P(\eta > X_{(n-k+1)}).$$

$$P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1} P(j \text{ observations are greater than } \eta).$$

$$P(\eta < X_{(k)}) = \sum_{j=0}^{k-1} P(j \text{ observations are less than } \eta).$$

By definition of the median we have.

$$P(X_i > \eta) = P(X_i < \eta) = 1/2.$$

So

$$\text{Prob}\{\text{exactly } j \text{ observations} > \eta) = \binom{n}{j} \cdot \frac{1}{2^n}$$

and

$$\text{Prob}(\eta > X_{(n-k+1)}) = \frac{1}{2^n} \sum_{j=0}^{k-1} \binom{n}{j}.$$

$$= \text{Prob}(\eta < X_{(k)})$$

So

$$\text{Prob}(X_{(k)} \le \eta \le X_{(n-k+1)})$$

$$= 1 - \frac{1}{2^{n-1}} \sum_{j=0}^{k-1} \binom{n}{j}.$$

For $n = 26$ and Y binomial with $p = 1/2$

we have

k	$P(Y \le k)$
5	.0012
6	.0047
7	.0145
8	.0378
9	.0843.

with $k = 8$.

$$P(Y < 8) = .0145.$$

$$P(Y > 26-8+1) = P(Y > 19) = .0145.$$

so $(X_{(8)}, X_{(19)})$ is a $1-(2(.0145)) \approx .97$ confidence interval for the mean.

For the Platinum data this is

$$(134.8, 135.8).$$

The sample mean confidence interval was.

$$(135.32, 138.75)$$

But note the data may not be independent

Bootstrapping for variability of location estimates.

~~Box plots~~

Dispersion measures

sample standard deviation

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

IQR
inter quartile range.

= difference between 25th and 75th percentiles of the ordered data.

MAD = median absolute deviation from median

= median of $|X_i - \tilde{X}|$

where $\tilde{X}$ is the median.

For a normal distribution

$$\sigma = \frac{IQR}{1.35} = \frac{MAD}{.675}$$

for the platinum data.

$$s = 4.45$$

$$\frac{IQR}{1.35} = .96..$$

$$\frac{MAD}{.675} = .934.$$

Box plots

1. horizontal lines at median and upper and lower quartiles. joined by a "box".

2. vertical line from upper quartile to most extreme data point within 1.5 IQR of upper quartile

   similarly for the lower quartile.

3. each data point beyond these vertical lines is marked with an dot or asterisk