PSQF 4143: Section 4

Brandon LeBeau

Properties of the mean

- Deviations sum to 0: Deviations are defined as $X - \bar{X}$.
- Sum of squared deviations is least for the mean (least squares property)
- Graphically, mean is a balancing point.
- In skewed distributions, mean is in the direction of the skew.

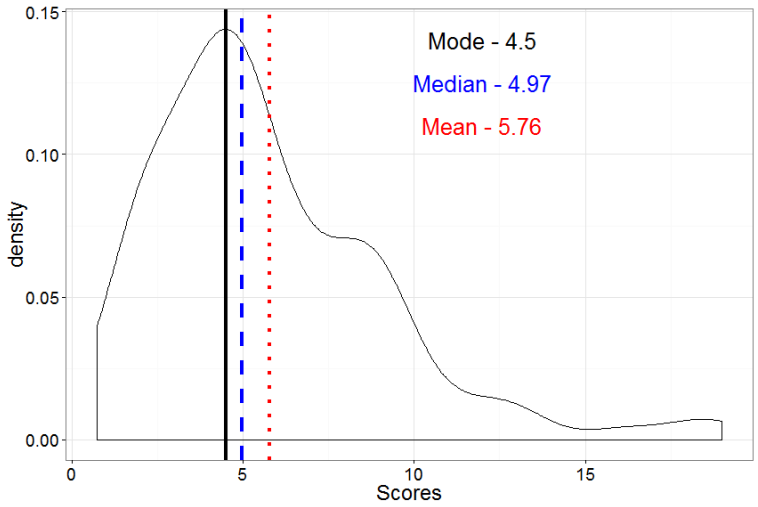


Figure 1: plot of chunk chisq

Variation Example

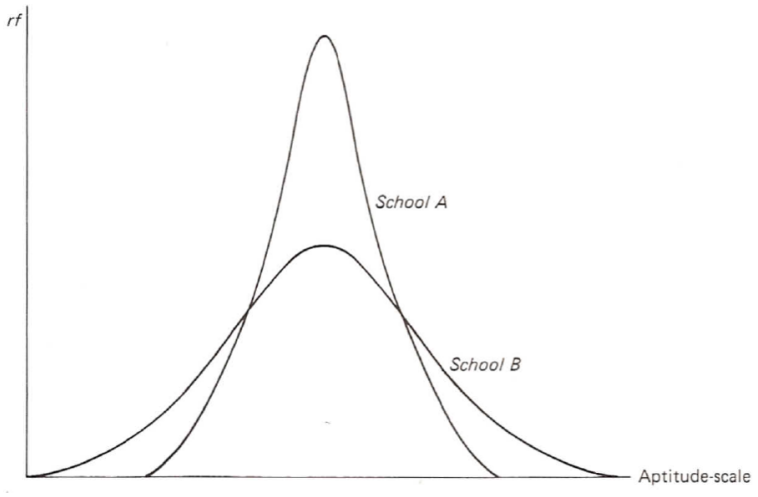


Figure 2: Variation

Variation

- Variability - how spread out scores are in a distribution
- Central Question of Research: Why are there differences on outcomes?
- Range:
  - Upper limit of highest score (H) minus low limit of lowest score (L)
  - If scores are continuous and rounded to the nearest point: $Range = H - L + 1$
  - If scores are discrete: $Range = H - L$
  - Using range as a verb: "ACT scores ranged from 17 to 24 for the group"
- Interquartile Range:
  - $IQR = Q_3 - Q_1$
- Is the range of the sample equal to the range of the population?
- Is the mean of a sample equal to the mean of the population?

Expected Value

- Expected value of a statistic is the mean over all possible samples; a long run average.

  - $E(samplemean) \quad populationmean(\mu)$
  - $E(samplerange) \quad populationrange$
- Range is very crude as it only takes into account two scores in our distribution.

Semi-Interquartile Range

$$Q = \frac{Q_3 - Q_1}{2}$$
$$Q = \frac{(Q_3 - Mdn) + (Mdn - Q_1)}{2}$$

- Q is the average of the distances from the median to Q_1 and Q_3.

Semi IQR Example

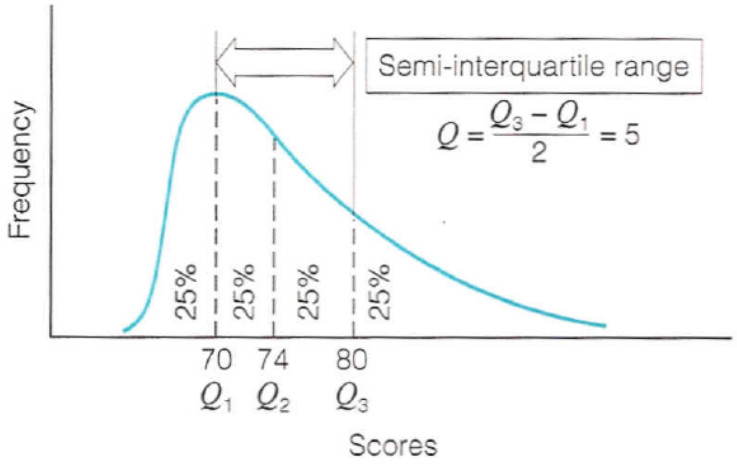


Figure 3: Semi IQR

- For non-symmetrical distributions, the semi IQR is interpreted as *roughly* half of the scores deviate by more than 5 units from the median and *roughly* half of the scores deviate by less than 5 units.
- For symmetrical distributions change *roughly* above to *exactly*.

Variability Measures with the Median

- Average deviation from the median:

$$\frac{\sum(X_i - Mdn)}{n}$$

  - Positive differences will tend to offset negative differences.
  - Half of the scores are above, half are below.

- Average absolute deviation from the median:

$$\frac{\sum |X_i - Mdn|}{n}$$

  - Uses all scores in the distribution
  - Influenced by extreme scores
  - Not too bad as an index of variability
  - One problem is that absolute values are difficult mathematically

Variability Measures with the Mean

- Average deviation from the mean:

$$\frac{\sum (X_i - \bar{X})}{n}$$

  - No Good, why?

- Average absolute deviation from the mean:

$$\frac{\sum |X_i - \bar{X}|}{n}$$

  - Not too bad as an index of variability
  - One problem is that absolute values are difficult mathematically
  - For more advanced statistical techniques, this will become a problem.

- Average squared deviations from the mean:

$$\frac{\sum (X_i - \bar{X})^2}{n}$$

  - This is formally called the variance
  - This is a very important statistical measure and the most used measure of variability.

Variance

- Sample Variance:

$$s^2 = \frac{\sum (X - \bar{X})^2}{n}$$

- Population Variance:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

  - Sum of Squares (SS) = $\sum(X-\mu)^2$
  - Mean Squares (MS) = variance or σ^2
- Interpreted as the average **squared** distance from the mean.
- Too bad the variance is interpreted in terms of squared units.

Example

X
3 4 5 6 7

Standard Deviation

- To convert away from squared units, simply take the square root of the variance.
- Sample SD:

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

- Population SD:

$$\sigma^2 = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X-\mu)^2}{n}}$$

- The SD is interpreted as the average deviation score from the mean.

Standard Deviation Properties

Figure 8. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within one SD of the average is shaded: 72% of the women differed from average by one SD (3 inches) or less.

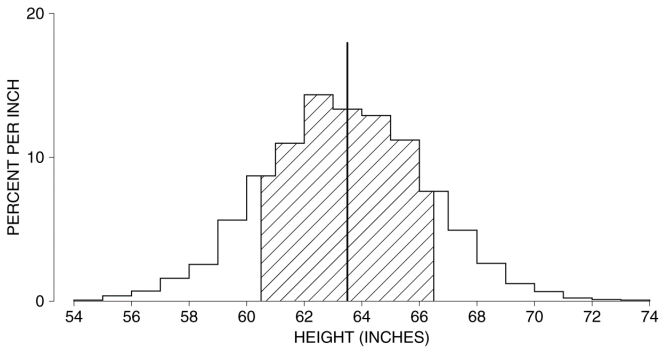
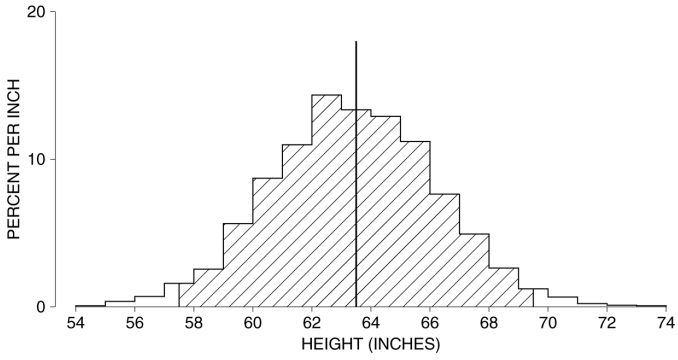


Figure 9. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within two SDs of the average is shaded: 97% of the women differed from average by two SDs (6 inches) or less.



Computational Formulas

- Variance:

$$s^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2 = \frac{\sum X_i^2}{n} - \bar{X}^2$$

- SD:

$$s = \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2} = \sqrt{\frac{\sum X_i^2}{n} - \bar{X}^2}$$

Calculate SD for Grouped Freq. Dist.

$$s = \sqrt{\frac{\sum f_j X_j^2}{n} - \left(\frac{\sum f_j X_j}{n}\right)^2} = \sqrt{\frac{\sum f_j X_j^2}{n} - \bar{X}^2}$$

- f_j is the frequency of interval j. - X_j is the midpoint of interval j.

Example

X	f_j	X_j	$f_j * X_j$
1 - 4	3	2.5	7.5
5 - 8	4	6.5	26
9 - 12	8	10.5	84
13 - 16	15	14.5	217.5
17 - 20	28	18.5	518
21 - 24	12	22.5	270
25 - 28	14	26.5	371
29 - 32	10	30.5	305
33 - 36	6	34.5	207

Relationship between Q and s

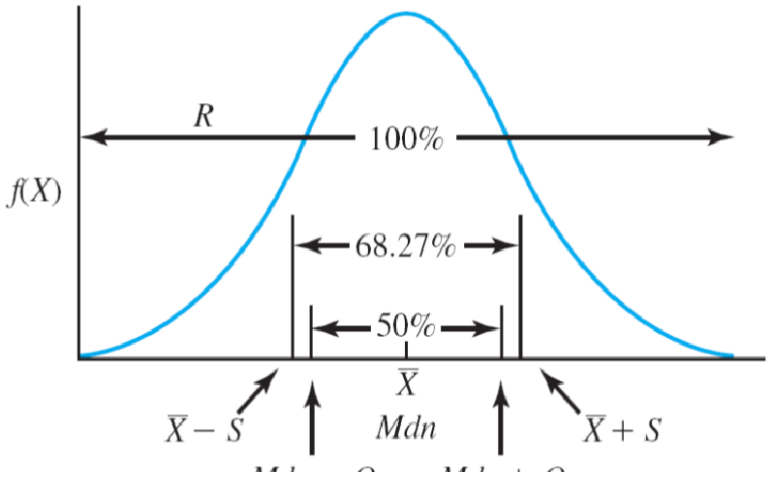
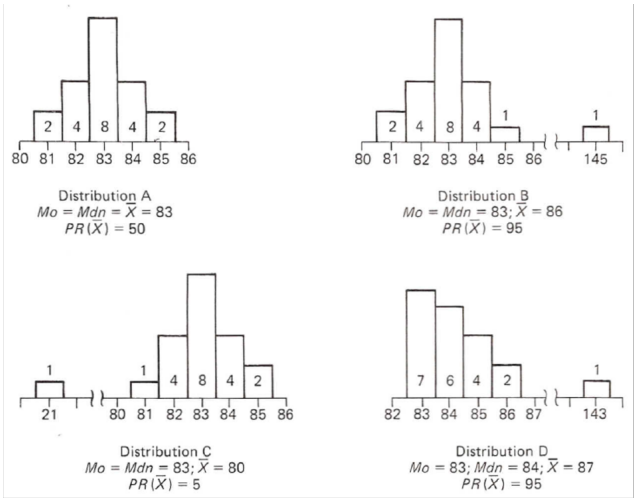


Figure 4: Q vs S

Q & s with skewed distributions



Distribution	S	Q
A. Unimodal, symmetrical	1.10	.75
B. One extreme score at right	13.57	.75
C. One extreme score at left	13.57	.75
D. J-shaped	12.88	.90

Figure 5: Q vs S

Income Example

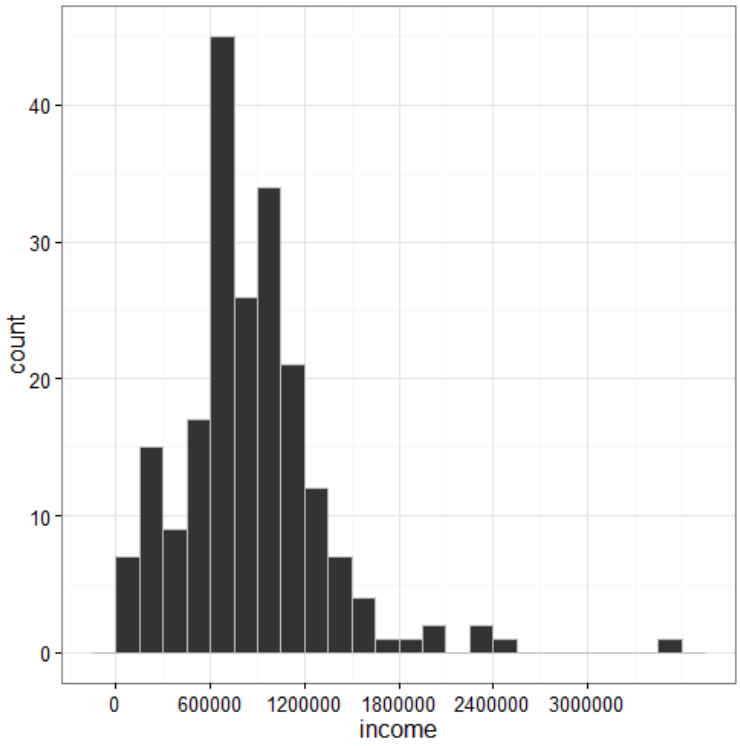


Figure 6: plot of chunk salary

Properties of Variability Measures

- Standard Deviation (Variance):
  - A distance measure
  - Preferred measure for symmetric quantitative variables
  - Commonly reported with the mean
  - Best sample stability
  - Widely used in advanced statistical procedures - mathematically tractable
  - Every score affects value
  - Fairly sensitive to extreme scores
  - Not appropriate for qualitative variables
- Semi-interquartile Range:
  - A distance measure
  - Commonly reported with the median
  - Sensitive to the number and not to the value of scores above or below Q3 and Q1 respectively

  - Appropriate for open-ended distributions
  - More sample fluctuation than SD
  - Rarely used in advanced statistical procedures - less mathematically tractable
- Range:
  - A distance measure
  - Simplest measure for quantitative variables
  - Highly influenced by sample fluctuations
  - Dependent on sample size
  - rarely used in advanced statistical procedures - less mathematically tractable

Measures of Variability Uses

1. Describe Distributions
2. Compare Distributions
3. Study the accuracy of certain measuring procedures
   - Consistency of raters rating figure skaters.
   - Estimate the population mean IQ of children aged 3 to 15.

Describing Distributions

- Used in conjunction with a measure of central tendency, variability helps us understand a distribution of scores.
  - Examples:
    * Consider a unimodal symmetrical distribution:
    * $Mdn = 50$, $Q = 10$.
    * $Mean = 38$, $SD = 6$.

Comparing Distributions

- If two distributions have the same score scale, then a direct comparison of variability is possible.
- Cannot compare when scales are different:
  - Example:
    * SAT: ranges from 200 to 800, $SD_{SAT} = 100$
    * ACT: ranges from 1 to 36, $SD_{ACT} = 5$

Accuracy of Measuring Procedures

- Roughly 60 million children in the US. aged 3 to 15.
- Not practical to measure the IQ of every child
- Settle for taking a sample of 1000 children and measure the IQ of each
- Use the sample mean as an estimate of the population mean
- How good is the estimate?
  - More complicated because the population mean is unknown

Population

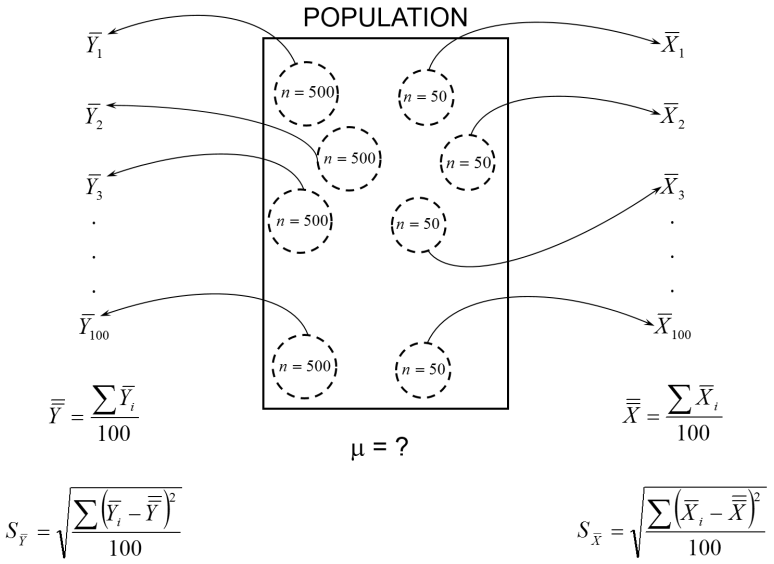


Figure 7: Popsamp

Standard Deviation 3

- Problem:
  - If we use the formula for the population on the sample, the variance for the sample is slightly too small - samples are closer to the sample mean than to the population mean. Recall that sum of squared deviations was smallest for the sample mean.
- Solution:
  - Instead of dividing by n, divide by $n - 1$.

Standard Deviations for Samples

- Variance:
$$s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$
- Standard Deviation:
$$s = \sqrt{s^2} = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

Standard Deviations for Samples 2

- By using $n-1$ instead of n, the expected value of the sample variance is the population variance. More formally:
$$E(s^2) = \sigma^2$$
- In words, the long run average from every possible sample from the population would now equal the population variance.

Index of Dispersion

$$D = \frac{c\left(n^2 - \sum n_j^2\right)}{n^2(c-1)}$$

- c is the number of categories
- n is the total number of observations
- n_j is the number of observations in category j

Index of Dispersion Example

TABLE 4.2-4 Marital Happiness Ratings of Women with Either a High School or a College Education

(i) Data

Rating	n_j, *High School Graduate*	n_j, *College Graduate*
Very happy	15	12
Moderately happy	28	39
Neutral	16	30
Unhappy	13	12
Very unhappy	8	3
	$n = 80$	$n = 96$
	Mo = Moderately happy	Mo = Moderately happy
	$D_{HG} = .96$	$D_{CG} = .88$

(ii) Computation of D

$$D = \frac{c\left(n^2 - \sum_{j=1}^{c} n_j^2\right)}{n^2(c-1)}$$

$$D_{HG} = \frac{5[(80)^2 - (15)^2 - (28)^2 - (16)^2 - (13)^2 - (8)^2]}{(80)^2(5-1)} = \frac{24{,}510}{25{,}600} = .96$$

$$D_{CG} = \frac{5[(96)^2 - (12)^2 - (39)^2 - (30)^2 - (12)^2 - (3)^2]}{(96)^2(5-1)} = \frac{32{,}490}{36{,}864} = .88$$

Figure 8: Dispersion

Skewness

$$Sk = \frac{\sum \left(X_i - \bar{X}\right)^3}{ns_X^3}$$

- $Sk = 0$, symmetrical - $Sk > 0$, positively skewed - $Sk < 0$, negatively skewed

Kurtosis

$$Kur = \frac{\sum \left(X_i - \bar{X}\right)^4}{ns_X^4} - 3$$

- $Kur = 0$, mesokurtic - $Kur > 0$, leptokurtic - $Kur < 0$, platykurtic

Distribution Differences

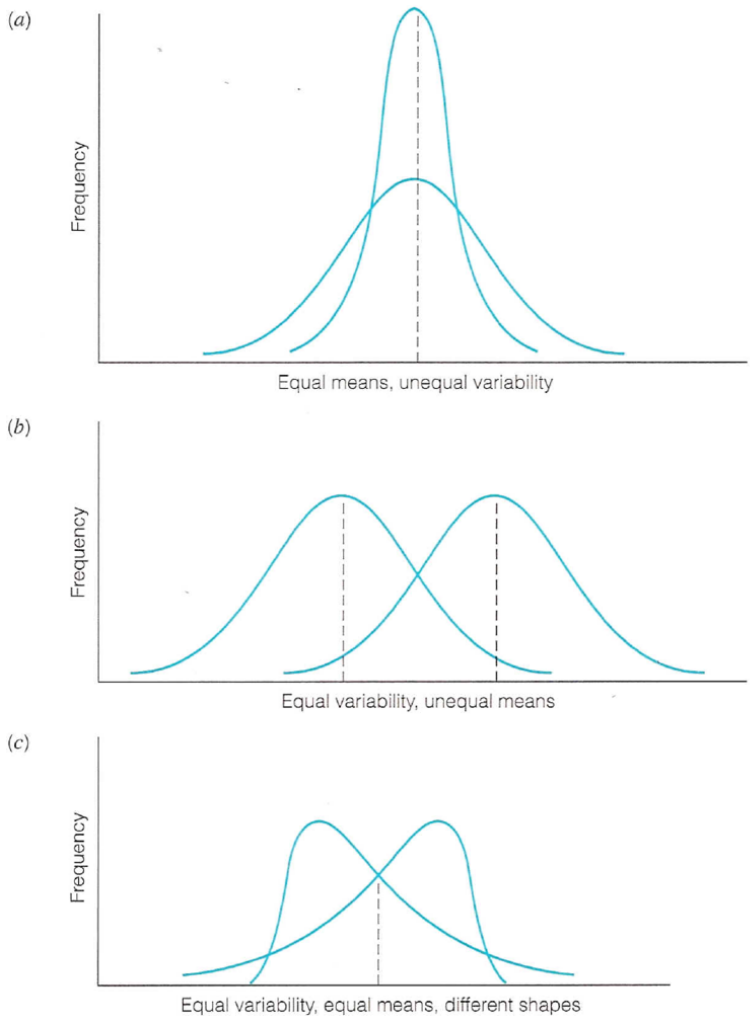


Figure 9: Dist Diff

Boxplot

- These are sometimes called a box and whisker plot

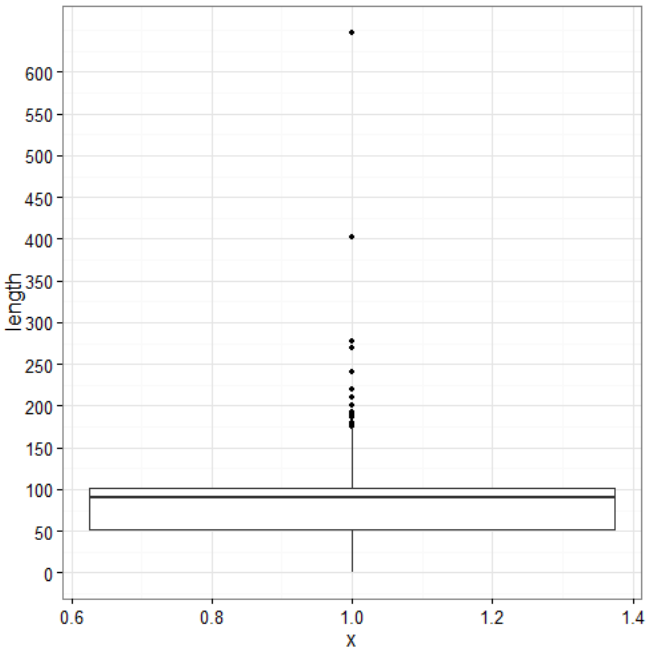


Figure 10: plot of chunk boxplot

Boxplots by group

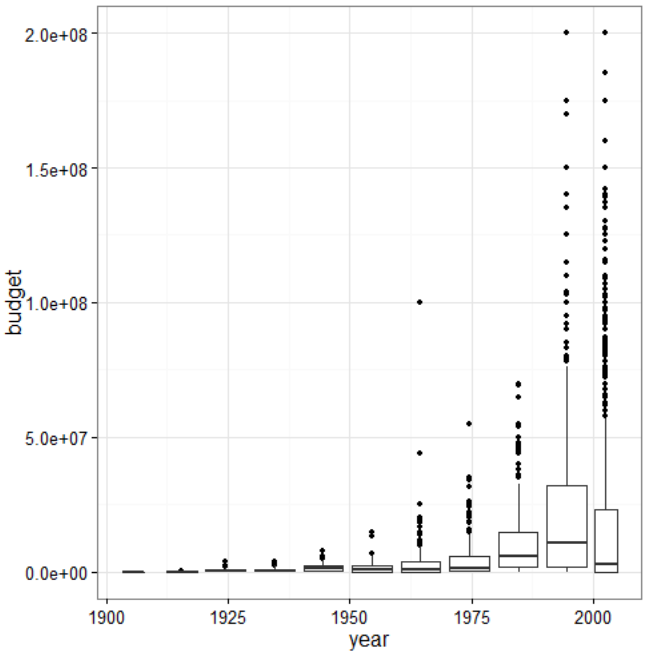


Figure 11: plot of chunk boxgroup