Exploratory Data Analysis

Week 3 Lectures: Letter Values, Box Plots & Letter Value Plots (UREDA Ch2&3)

David B King, Ph.D.

September 15, 2015

Stem-and-leaf display

- Informative for $n < 200$
- For large data sets, stem-and-leaf displays not useful
- n large, stem-and-leaf display is too crowded
- Large data sets carry more information, but also require more summarization

Summary statistics

- Classical statistics

$$\text{Sample mean: } \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Exploratory data analysis:

  Use summaries based on sorting and counting

Summary statistics

- Classical statistics

$$\text{Sample mean: } \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Exploratory data analysis:

  Use summaries based on sorting and counting – resistant

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- Sort data $x_1, x_2, \ldots, x_n$ into ascending order

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- Sort data $x_1, x_2, \ldots, x_n$ into ascending order
- Rank

$$\text{upward rank} + \text{downward rank} = n + 1$$

- Depth $= \min\{\text{upward rank, downward rank}\}$

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- Sort data $x_1, x_2, \ldots, x_n$ into ascending order
- Rank

$$\text{upward rank} + \text{downward rank} = n + 1$$

- Depth $= \min\{\text{upward rank, downward rank}\}$
  - Both $x_{(2)}$ and $x_{(n-1)}$ have depth 2.
  - The depth of $x_{(i)}$ is the smaller of i and $n + 1 - i$.
- Median: center of the ordered sample
  - depth $= \dfrac{n+1}{2}$
  - If $n = 2k$, median $= \dfrac{1}{2}(x_{(k)} + x_{(k+1)})$

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- Sort data $x_1, x_2, \ldots, x_n$ into ascending order
- Rank

$$\text{upward rank} + \text{downward rank} = n + 1$$

- Depth $= \min\{\text{upward rank, downward rank}\}$
  - Both $x_{(2)}$ and $x_{(n-1)}$ have depth 2.
  - The depth of $x_{(i)}$ is the smaller of i and $n + 1 - i$.
- Median: center of the ordered sample
  - depth $= \dfrac{n+1}{2}$
  - If $n = 2k$, median $= \dfrac{1}{2}(x_{(k)} + x_{(k+1)})$
- Extremes: $x_{(1)}$, $x_{(n)}$, both with depth 1

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- Sort data $x_1, x_2, \ldots, x_n$ into ascending order
- Rank

$$\text{upward rank} + \text{downward rank} = n + 1$$

- Depth $= \min\{\text{upward rank, downward rank}\}$
  - Both $x_{(2)}$ and $x_{(n-1)}$ have depth 2.
  - The depth of $x_{(i)}$ is the smaller of i and $n + 1 - i$.
- Median: center of the ordered sample
  - depth $= \dfrac{n+1}{2}$
  - If $n = 2k$, median $= \dfrac{1}{2}(x_{(k)} + x_{(k+1)})$
- Extremes: $x_{(1)}$, $x_{(n)}$, both with depth 1
- Hinges (a form of quantiles) or the fourths

$$\text{depth of fourth} = \frac{[\text{depth of median}] + 1}{2}$$

where $[x]$ represents largest integer not greater than x.

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Examples

- $n = 9$
  - extremes: $x_{(1)}$, $x_{(9)}$
  - median: depth $= \dfrac{n+1}{2} = 5$, $\quad x_{(5)}$
  - fourths:

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Examples

- $n = 9$
  - extremes: $x_{(1)}$, $x_{(9)}$
  - median: depth $= \dfrac{n+1}{2} = 5$, $x_{(5)}$
  - fourths: depth $= \dfrac{5+1}{2} = 3$, $x_{(3)}$, $x_{(7)}$
- $n = 10$
  - extremes: $x_{(1)}$, $x_{(10)}$
  - median: depth $= \dfrac{n+1}{2} = 5.5$,

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Examples

- $n = 9$
  - extremes: $x_{(1)}$, $x_{(9)}$
  - median: depth $= \dfrac{n+1}{2} = 5$, $\quad x_{(5)}$
  - fourths: depth $= \dfrac{5+1}{2} = 3$, $\quad x_{(3)}$, $x_{(7)}$
- $n = 10$
  - extremes: $x_{(1)}$, $x_{(10)}$
  - median: depth $= \dfrac{n+1}{2} = 5.5$, $\quad \dfrac{x_{(5)} + x_{(6)}}{2}$

Order statistics: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Examples

- $n = 9$
  - extremes: $x_{(1)}$, $x_{(9)}$
  - median: depth $= \dfrac{n+1}{2} = 5$, $\quad x_{(5)}$
  - fourths: depth $= \dfrac{5+1}{2} = 3$, $\quad x_{(3)}$, $x_{(7)}$
- $n = 10$
  - extremes: $x_{(1)}$, $x_{(10)}$
  - median: depth $= \dfrac{n+1}{2} = 5.5$, $\quad \dfrac{x_{(5)} + x_{(6)}}{2}$
  - fourths: depth $= \dfrac{5+1}{2} = 3$, $\quad x_{(3)}$, $x_{(8)}$

Letter values

- 5-number summary:

  median, fourths, extremes

- 7-number summary:

  5-number summary PLUS eighths

$$\text{depth of eighth} = \frac{[\text{depth of fourth}] + 1}{2}$$

- When batches get larger, can include more summary values.

$$d_L = \frac{[\text{previous depth}] + 1}{2}$$

- When d_L is half-integer, LV = avg 2 adjacent order statistics
- Except for the median, letter values come in paris: a lower one and up upper one.

Letter values: tags
1-letter tags

Tags	Tail areas for continuous distributions
1: extremes	
M: median	$1/2$
F: fourths	$1/4$
E: eighths	$1/8$
D	$1/16$
C	$1/32$
$\vdots$	$\vdots$

- Estimate quantiles corresponding to tail areas 2^{-k}
- Actual tail area is closer to $\dfrac{d_L - 1/3}{n + 1/3}$

Letter values as measures

- Location summary: median, trimean

  $$\text{trimean} = \frac{1}{4}(\text{lower fourth}) + \frac{1}{2}(\text{median}) + \frac{1}{4}(\text{upper fourth})$$

Letter values as measures

- Location summary: median, trimean

  $$\text{trimean} = \frac{1}{4}(\text{lower fourth}) + \frac{1}{2}(\text{median}) + \frac{1}{4}(\text{upper fourth})$$

- Spread summary:
  – fourth-spread or F-spread, d_F

  $$d_F = \text{(upper fourth) - (lower fourth)}$$

Letter values as measures

- Location summary: median, trimean

  $$\text{trimean} = \tfrac{1}{4}(\text{lower fourth}) + \tfrac{1}{2}(\text{median}) + \tfrac{1}{4}(\text{upper fourth})$$

- Spread summary:
  - fourth-spread or F-spread, d_F

    $$d_F = \text{(upper fourth) - (lower fourth)}$$

  - range

Letter values as measures

- Location summary: median, trimean

  $$\text{trimean} = \frac{1}{4}(\text{lower fourth}) + \frac{1}{2}(\text{median}) + \frac{1}{4}(\text{upper fourth})$$

- Spread summary:
  - fourth-spread or F-spread, d_F

    $$d_F = \text{(upper fourth) - (lower fourth)}$$

  - range: difference between extremes
  - Which one is resistant?
- Compare batches: F-spread and median help to choose a scale of measurement.
- Outliers and outside values

Letter values as measures

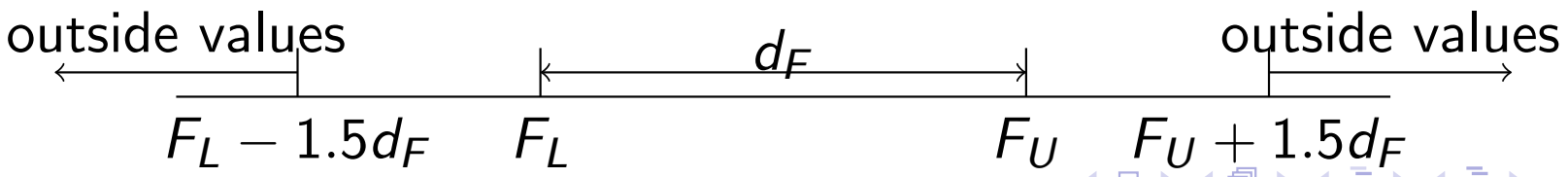
- Location summary: median, trimean

  $$\text{trimean} = \frac{1}{4}(\text{lower fourth}) + \frac{1}{2}(\text{median}) + \frac{1}{4}(\text{upper fourth})$$

- Spread summary:
  - fourth-spread or F-spread, d_F

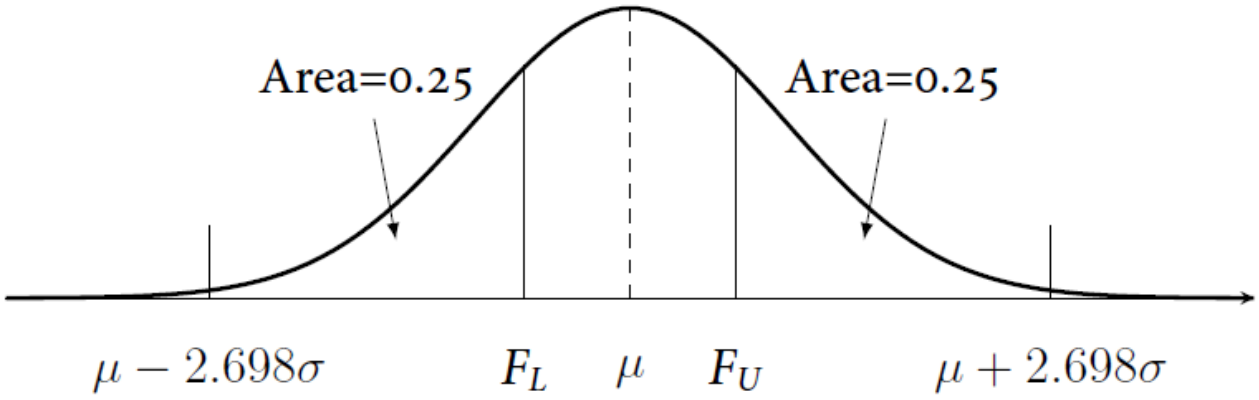
    $$d_F = \text{(upper fourth) - (lower fourth)}$$

  - range: difference between extremes
  - Which one is resistant?
- Compare batches: F-spread and median help to choose a scale of measurement.
- Outliers and outside values
  Rule of thumb:



Letter values: Gaussian distribution

Example: $N(\mu, \sigma^2)$



- Median = μ,
- Fourths: $F_u = \mu + 0.6745\sigma$, $F_L = \mu - 0.6745\sigma$
- F-spread: $d_F = F_u - F_L = 1.349\sigma$
- Outside cutoffs:

$$F_u + 1.5d_F = \mu + 2.698\sigma, \qquad F_L - 1.5d_F = \mu - 2.698\sigma$$

- Outside area = 0.00698

Letter values: Gaussian distribution

Example (continued): $N(\mu, \sigma^2)$

- In finite samples, the average fraction of observations beyond the "outside" cutoffs is substantially larger than the population value.
- Average number "outside" for a single batch

$$0.4 + 0.007n$$

(From a simulation study by Hoaglin, Iglewics, and Tukey, 1981)

Letter values: more resistant measures

Replacements for standard deviation or variance

- F-spread v.s. Sample standard deviation s

Letter values: more resistant measures

Replacements for standard deviation or variance

- F-spread v.s. Sample standard deviation s
- Gaussian distribution

$$d_F = 1.349\sigma \Rightarrow \sigma = \frac{d_F}{1.349}$$

- F-pseudosigma: estimate of σ

$$\frac{\text{data F-spread}}{1.349}$$

- F-pseudovariance: estimate of σ^2

$$\left(\frac{\text{data F-spread}}{1.349}\right)^2$$

Letter values: more resistant measures

Replacements for standard deviation or variance

- F-spread v.s. Sample standard deviation s
- Gaussian distribution

$$d_F = 1.349\sigma \Rightarrow \sigma = \frac{d_F}{1.349}$$

- F-pseudosigma: estimate of σ

$$\frac{\text{data F-spread}}{1.349}$$

- F-pseudovariance: estimate of σ^2

$$\left(\frac{\text{data F-spread}}{1.349}\right)^2$$

- Use other letter values:

$$\frac{\text{data letter-spread}}{\text{standard Gaussian value}}$$

Letter values: display

Exercise 1: $n = 65$

28 33 36 36 37 37 38 38 39 39 40 41 42 43 44 44
46 46 47 47 47 47 47 47 48 48 48 48 48 49 49 49
49 50 50 50 51 51 52 52 52 53 54 55 55 55 56 56
57 57 57 57 58 59 60 60 61 62 65 65 67 68 68 71 73

Letter value program in R

```
lval <- function(x) {
 #tag <- c("M ","F ","E ","D ","C ","B ","A ","Z ","Y ","X ","W ","V","U","T",
 # "S","R","Q","P","O","N")
 # gau <- abs(qnorm(c(.25,.125,1/16,1/32,1/64,1/128,1/256,1/512,1/1024,1/2048,
 #    1/4096, 1/8192, 1/16384, 1/32768, 1/65536)))
  tag <- c("M",LETTERS[6:1],LETTERS[26:14])

  gau <- abs(qnorm(1/2^(2:20)))

# col 1 = depth; 2 = lower; 3 = upper; 4 = mid; 5 = spread; 6 = pseudo-s

  y <- sort(x[!is.na(x)])
  n <- length(y)
  m <- ceiling(log(n)/log(2)) + 1
  depth    <- rep(0,m)
  depth[1] <- (1 + n)/2

  for (j in 2:m) {depth[j] <- (1 + floor(depth[j-1]))/2 }
```

See the attached R code.

Letter value program in R

```
  ndepth <- n+1 - depth
  out <- matrix(0, m, 6)
  dimnames(out) <- list(tag[1:m],
          c("Depth", "Lower","Upper", "Mid", "Spread","pseudo-s"))
  out[1,2:3] <- median(y)
  out[,1] <- depth


  for (k in 2:m) {
    out[k,2] <- ifelse(depth[k] - round(depth[k]) == 0,
                y[depth[k]], (y[depth[k]-.5]+y[depth[k]+.5])/2 )
    out[k,3] <- ifelse(ndepth[k] - round(ndepth[k]) == 0,
                y[ndepth[k]], (y[ndepth[k]-.5]+y[ndepth[k]+.5])/2 )
  }

  out[1:m,4] <- (out[1:m,2] + out[1:m,3])/2
  out[2:m,5] <- out[2:m,3] - out[2:m,2]
  out[2:m,6] <- out[2:m,5]/(2*gau[1:(m-1)])
  round(out,4)
}
```

Letter values: display

Letter-value display:

```
> data1 <- scan()
> lval(data1)
  Depth Lower Upper   Mid Spread pseudo-s
M  33.0  49.0    49 49.00    0.0   0.0000
F  17.0  46.0    57 51.50   11.0   8.1543
E   9.0  39.0    61 50.00   22.0   9.5623
D   5.0  37.0    67 52.00   30.0   9.7776
C   3.0  36.0    68 52.00   32.0   8.5895
B   2.0  33.0    71 52.00   38.0   8.8213
A   1.5  30.5    72 51.25   41.5   8.5830
Z   1.0  28.0    73 50.50   45.0   8.4584
```

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be equal to the median.

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be equal to the median.
- If the data were skewed to the right, the midsummaries would

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be equal to the median.
- If the data were skewed to the right, the midsummaries would increase as they came from letter values further into the tails.

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be equal to the median.
- If the data were skewed to the right, the midsummaries would increase as they came from letter values further into the tails.
- For data skewed to the left, they would

Midsummary

- Midsummaries ("mids" for short)
  – Define a set of midsummaries: for each pair of letter values, the corresponding midsummary is the average of the two letter values.
- Using the full set of midsummaries provides more resistance.
- In a perfectly symmetric batch, all midsummaries would be equal to the median.
- If the data were skewed to the right, the midsummaries would increase as they came from letter values further into the tails.
- For data skewed to the left, they would decrease.

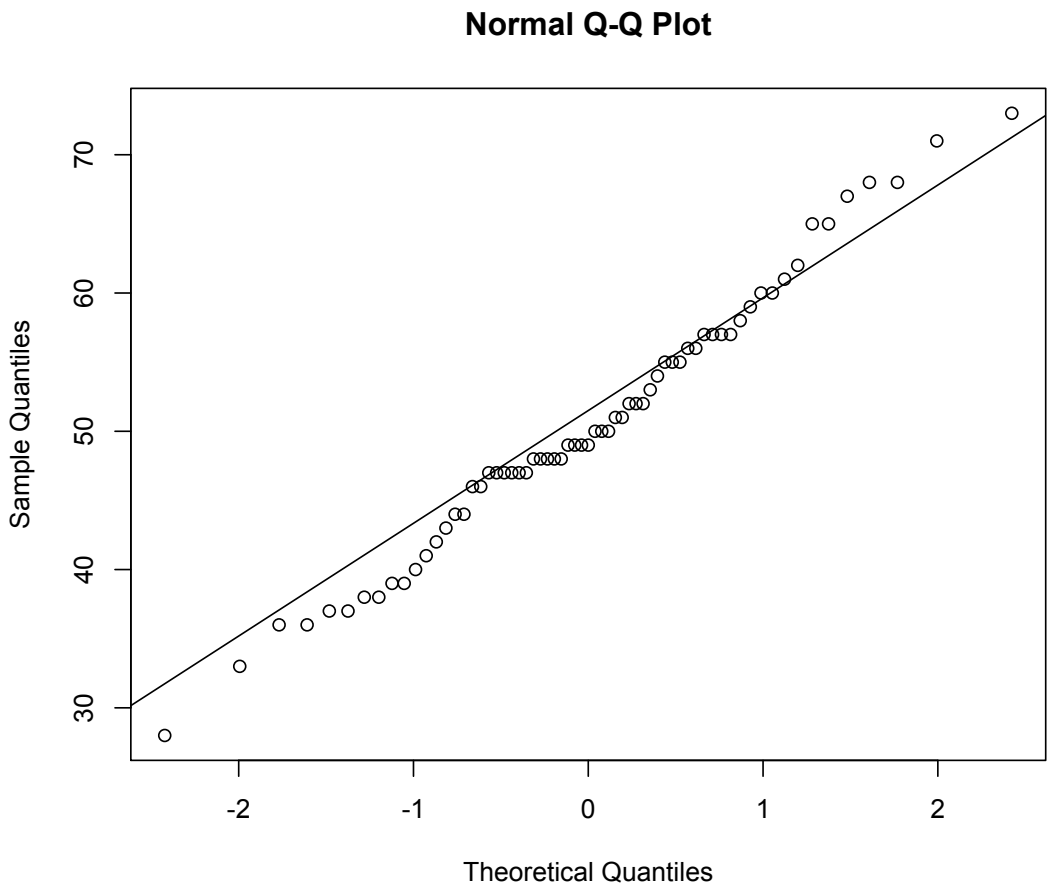
Letter values: display

Stem-and-leaf display:

```
> stem(data1)

  The decimal point is 1 digit(s) to the right of the |

  2 | 8
  3 | 3
  3 | 66778899
  4 | 012344
  4 | 66777777888889999
  5 | 0001122234
  5 | 55566777789
  6 | 0012
  6 | 55788
  7 | 13
```

Letter values: display

```
> qqnorm(data1)
> qqline(data1)
```



Letter values: display

Exercise 2: $n = 65$

13	18	19	21	28	32	33	33	38	40	42	46	55
57	59	67	73	74	76	78	85	97	101	102	106	107
113	113	120	120	124	125	125	127	128	129	135	138	149
168	168	183	184	193	204	205	228	231	233	240	241	260
274	275	286	312	320	334	337	361	467	486	711	743	759

Letter values: display

Letter-value display:

```
> data2 <- scan()
> lval(data2)
  Depth Lower Upper    Mid Spread pseudo-s
M  33.0 125.0   125 125.00    0.0   0.0000
F  17.0  73.0   233 153.00  160.0 118.6082
E   9.0  38.0   320 179.00  282.0 122.5715
D   5.0  28.0   467 247.50  439.0 143.0787
C   3.0  19.0   711 365.00  692.0 185.7487
B   2.0  18.0   743 380.50  725.0 168.3013
A   1.5  15.5   751 383.25  735.5 152.1162
Z   1.0  13.0   759 386.00  746.0 140.2220
```

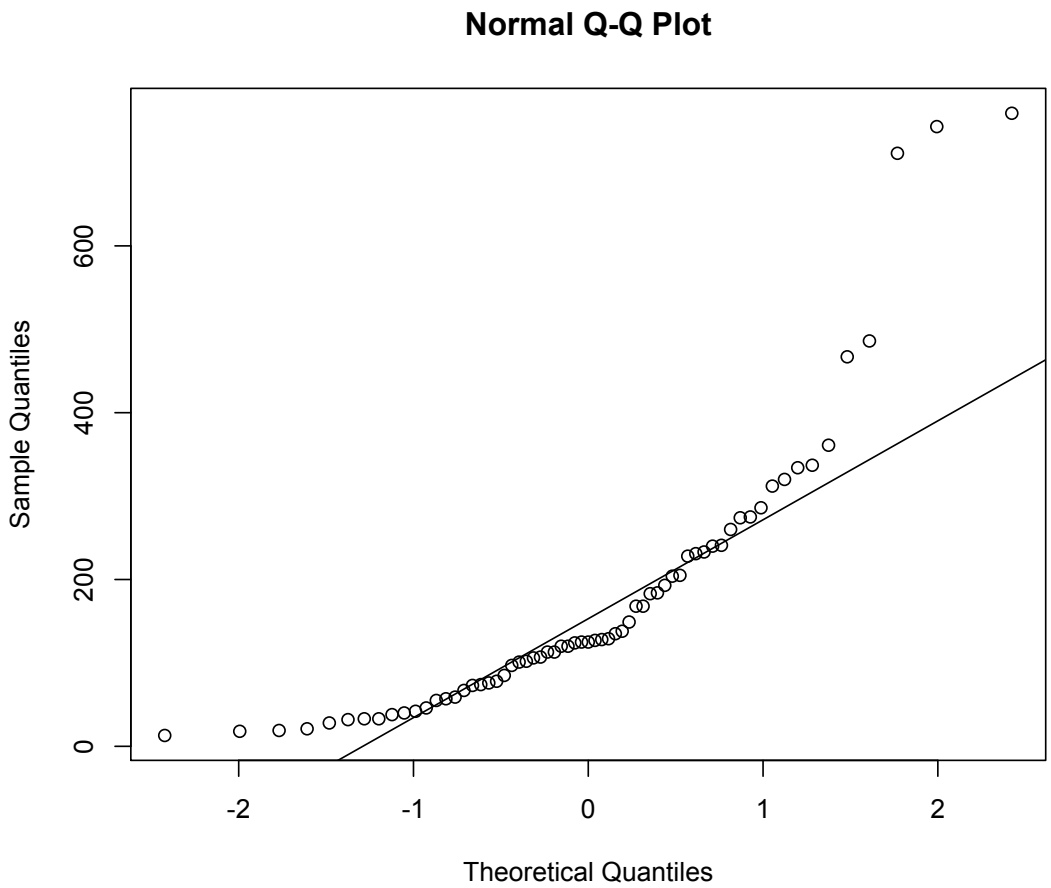
Letter values: display

Stem-and-leaf display:

```
> stem(data2)

The decimal point is 2 digit(s) to the right of the |

  0 | 1222333344566677889
  1 | 000111122233333344577889
  2 | 01333446789
  3 | 12346
  4 | 79
  5 |
  6 |
  7 | 146
```

Letter values: display

```
> qqnorm(data2)
> qqline(data2)
```



Letter values: Overview

1. n large $\Rightarrow$ stem-and-leaf display is too crowded
2. For population or sample with single mode, LV display is good summary of data distribution, especially in tails (multimodal: use smoothed histogram or nonparametric density estimate)

Letter values: Overview

1. n large $\Rightarrow$ stem-and-leaf display is too crowded
2. For population or sample with single mode, LV display is good summary of data distribution, especially in tails (multimodal: use smoothed histogram or nonparametric density estimate)
3. LVs are either a data value or average of 2 adjacent data values (simple for hand calculation)
4. LVs correspond *roughly* to quantiles with tail areas 2^{-j}
5. LVs, defined in terms of their depths,

$$d_j \equiv \text{depth}(LV_j) = \frac{1 + [d_{j-1}]}{2}$$

is about the most sensible way to achieve this, in that

$$P\{X \leq LV_j\} \approx 2^{-j}$$

Conventional depths v.s. Ideal depths

- More precisely, cdf of X is

$$F_X(LV_j) \equiv P\{X \leq LV_j\} \approx \frac{d_j - \frac{1}{3}}{n + \frac{1}{3}}$$

$$\Rightarrow LV_j \approx F_X^{-1}\left(\frac{d_j - 1/3}{n + 1/3}\right),$$

  not $F^{-1}(2^{-j})$ or $F^{-1}(1 - 2^{-j})$.

- Ideal depth

$$\text{depth} = \left(n + \frac{1}{3}\right) \times (\text{tail area}) + \frac{1}{3}$$

- Why do we use conventional depths?
  - Conventional depths: always deeper into the batch, but difference is less than one unit
  - Ideal depths: complex fractions, not in hand calculation
  - Ideal depths lose resistance when n is small
  - Ideal depths: little gain in bias and variance

More theory on ideal letter values

1. Blom (1958): A family of definitions for "the fraction of the data to the left of any spacified point x", parametrized by α

$$(\text{fraction} \leq x_{(i)}) = \frac{i - \alpha}{n + 1 - 2\alpha}$$

   - $\alpha = \frac{1}{2} \Rightarrow \frac{i-1/2}{n}$: Simple fraction
   - $\alpha = 0 \Rightarrow \frac{i}{n+1}$: Intervals of equal probability
   - $\alpha = \frac{1}{3} \Rightarrow \frac{i-\frac{1}{3}}{n+\frac{1}{3}}$, our choice of depths for LVs
   - The median of the distribution of $X_{(i)}$ in a sample of n is, very closely, at the point where the value of the cumulative distribution function equals $(i - \frac{1}{3})/(n + \frac{1}{3})$, i.e.

$$med(X_{(i)}) \approx F_X^{-1}\left(\frac{i - \frac{1}{3}}{n + \frac{1}{3}}\right)$$

Distribution of $X_{(i)}$

Recall the probability density function (PDF) of $X_{(i)}$ is given by

$$gx_{(i)}(x) = \frac{n!}{(i-1)!(n-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i} f(x).$$

Thus the median of $X_{(i)}$, denoted $x_{\{i,0.5\}}$ is the point of the x-axis such that

$$\frac{n!}{(i-1)!(n-i)!}\int_{-\infty}^{x_{\{i,0.5\}}}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i} f(x)dx = \frac{1}{2}.$$

Since $dF(x) = f(x)dx$ we can write the integral above as

$$\frac{n!}{(i-1)!(n-i)!}\int_{-\infty}^{F_{\{i,0.5\}}}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i} dF = \frac{1}{2}.$$

with $F_{\{i,0.5\}} = F(x_{\{i,0.5\}})$.

The Beta Function and Incomplete Beta Function

In mathematics the **Beta Function** $B(\alpha, \beta)$ is defined as

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The **Incomplete Beta Function** is defined as

$$B(u, \alpha, \beta) = \int_0^u t^{\alpha-1}(1-t)^{\beta-1}dt$$

The **Regularized Incomplete Beta Function** is defined as

$$I(u, \alpha, \beta) = \frac{B(u, \alpha, \beta)}{B(\alpha, \beta)}$$

The Median of $X_{(i)}$

From the equation for $F_{\{i,0.5\}}$ we see that

$$\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} B(F_{\{i,0.5\}}, i, n-i+1) = \frac{B(F_{\{i,0.5\}}, i, n-i+1)}{B(i, n-i+1)}$$

$$= I(F_{\{i,0.5\}}, i, n-i+1) = \frac{1}{2}.$$

The above suggests that

$$F_{\{i,0.5\}} = I^{-1}(1/2, i, n-i+1)$$

Thus,

$$x_{\{i,0.5\}} = F^{-1}(I^{-1}(1/2, i, n-i+1)) \approx F^{-1}\left(\frac{i+1/3}{(n+1/3)}\right).$$

1. Blom (1958) showed that a good numeric approximation for $I^{-1}(1/2, i, n-i+1)$ is $\frac{i-1/3}{n+1/3}$.
2. Approximation faces it's toughest test for small n and extreme i.

More theory on ideal letter values

2. How close? For n = 33 (all LVs are single data values):

j	Tag	d_j	ideal	D_j	dif	%dif
1	M	17	1/2	0.50	0.00	0%
2	F	9	1/4	0.26	0.01	4%
3	E	5	1/8	0.14	0.015	12%
4	D	3	1/16	0.08	0.0175	28%
5	C	2	1/32	0.05	0.01875	60%
6	B	1	1/64	0.02	0.004375	28%

   where $D_j = \frac{d_j - 1/3}{n + 1/3}$.
3. Approximation is close in terms of actual tail areas but relatively worse when $d_j \leq 5$ (extreme data values).

Spacing of letter values

When are the letter values equally spaced?

1. Logistic distribution:

$$F(x) = \frac{e^x}{1 + e^x}$$

   Differences in LVs nearly constant, $\log_e 2 = 0.69315$, from eighths on out (omitted proof, refer to UREDA).

2. Gaussian distribution (similar shape, but pdf goes to zero more rapidly):
   LVs trend steadily closer together as the tail area decreases.

How well do letter values work?

How well can we predict that value of an unselected order statistics by using the nearest selected order statistics?

1. Mosteller showed that as $n \to \infty$

$$\text{Corr}^2\left[X_{(i)}, X_{(j)}\right] \approx \frac{p/(1-p)}{q/(1-q)}$$

with $p = i/n \leq q = j/n$.

2. If we take $q = (1/2)^r$ and $p = (1/2)^{(r+1)}$ then $\text{Corr}^2\left[X_{(i)}, X_{(j)}\right] = \frac{(1/2)^{r+1}}{(1-(1/2)^{r+1})} \frac{(1-(1/2)^r)}{(1/2)^r} \to 1/2$ for large r. Asymptotic correlation between adjacent LVs approach ≈ 0.707.
3. Correlation between an unselected order statistics between median and the fourths is at least 0.76.
4. The LVs are a very effective set of selected order statistics. Little information in the ordered sample is lost when we use the LVs to summarize it.

Question for Class

If $X_1, X_2, \ldots, X_n$ are independent ...

Question for Class

If $X_1, X_2, \ldots, X_n$ are independent ...

why aren't $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$?

Question for Class

If $X_1, X_2, \ldots, X_n$ are independent ...

why aren't $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$?

Bivariate distribution of two ordered statistics:

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}[F_X(x)]^{j-1}$$
$$[F_X(y) - F_X(x)]^{k-1-j}[1 - F_X(y)]^{n-k} f_X(x) f_X(y)$$
$$\text{where } x \leq y$$

Question for Class

If $X_1, X_2, \ldots, X_n$ are independent ...

why aren't $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$?

Bivariate distribution of two ordered statistics:

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}[F_X(x)]^{j-1}$$
$$[F_X(y) - F_X(x)]^{k-1-j}[1 - F_X(y)]^{n-k} f_X(x) f_X(y)$$
$$\text{where } x \leq y$$

Joint distribution of all ordered statistics:

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! f_X(x_1) \cdots f_X(x_n)$$
$$\text{where } x_1 \leq x_2 \leq \cdots \leq x_n.$$

R code for letter value display

I have saved the letter value functions from Karen Kafadar in Canvas (see lvalprogs.r)

- lval()
- lval.sub()
- lvplot()

Graphical and Non-graphical Overview

	non-graphical	graphical
Batch of numbers	five-number display	stem-and-leaf
	letter value display	dotplot
		Boxplot

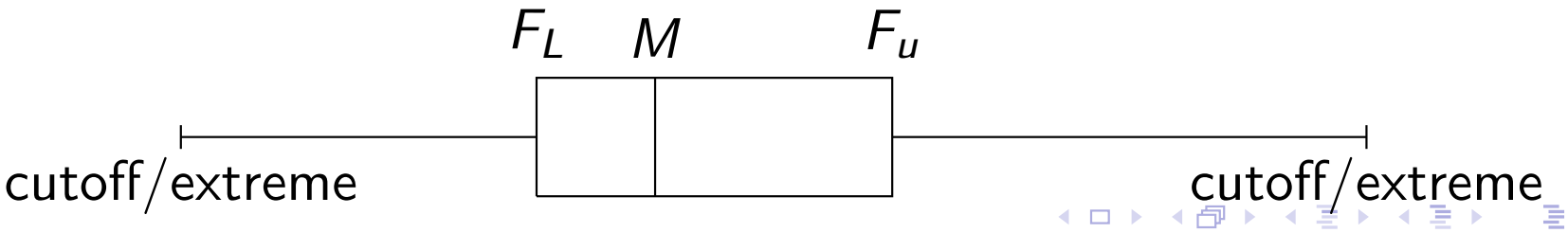
The Boxplot

This is shorted name for the original "box and whiskers plot".

- Display distributional shape of univariate data
- Focus on location, spread, skewness, tail length, outlying data points
- Compare batches: w/ parallel boxplots
- Initial concept: Tukey "schematic plot" (early EDA)
- Later enhancements: notches, bivariate (relplot; bagplot), LV boxplots

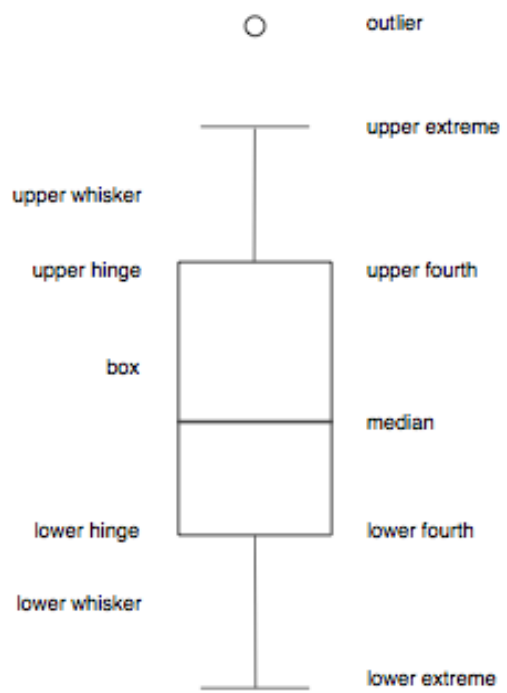
Construct a boxplot for one single batch

1. Construct 5-number summary (median, fourths, extremes)
2. Calculate F-spread
3. Step $= 1.5\times$ F-Spread
4. Observations beyond Inner Fences $=$ One step beyond fourths ("out")
5. Observations beyond Outer Fences $=$ Two steps beyond fourths ("far out")
6. Draw a box with:
   - Ends at fourths
   - Crossbar at the median
   - Draw a line from each end (tails/whiskers)
   - Stop whiskers at Observations just within inner fences

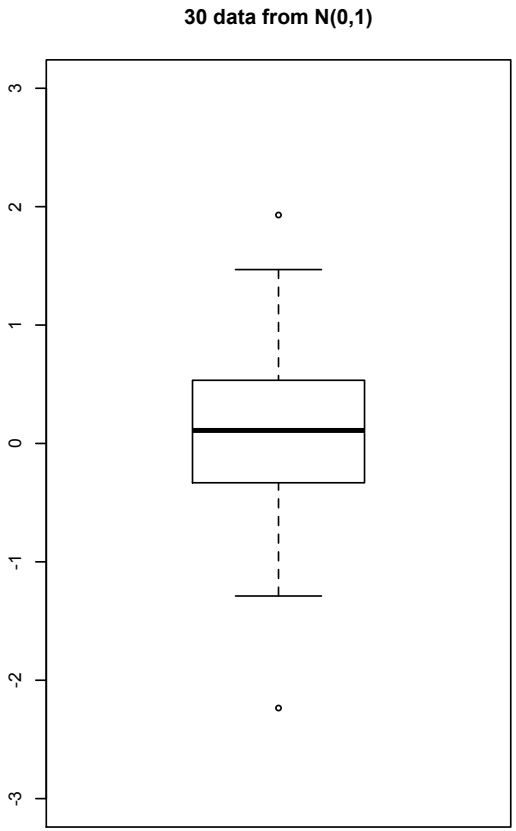


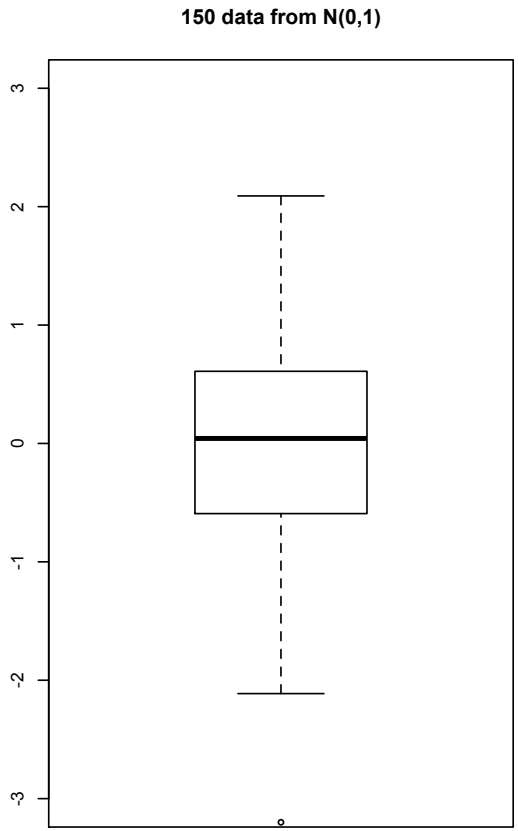
Construct a boxplot for one single batch

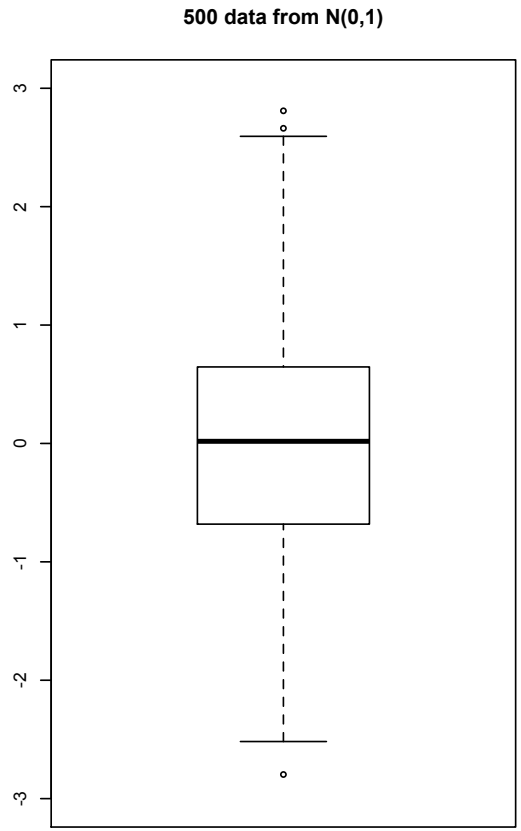
- Labels on the left: give names for graphic elements
- Labels on the right: give the corresponding summary statistics.
- R: *boxplot(data)* or *boxplot(data, horizontal=T)*



Boxplot Examples





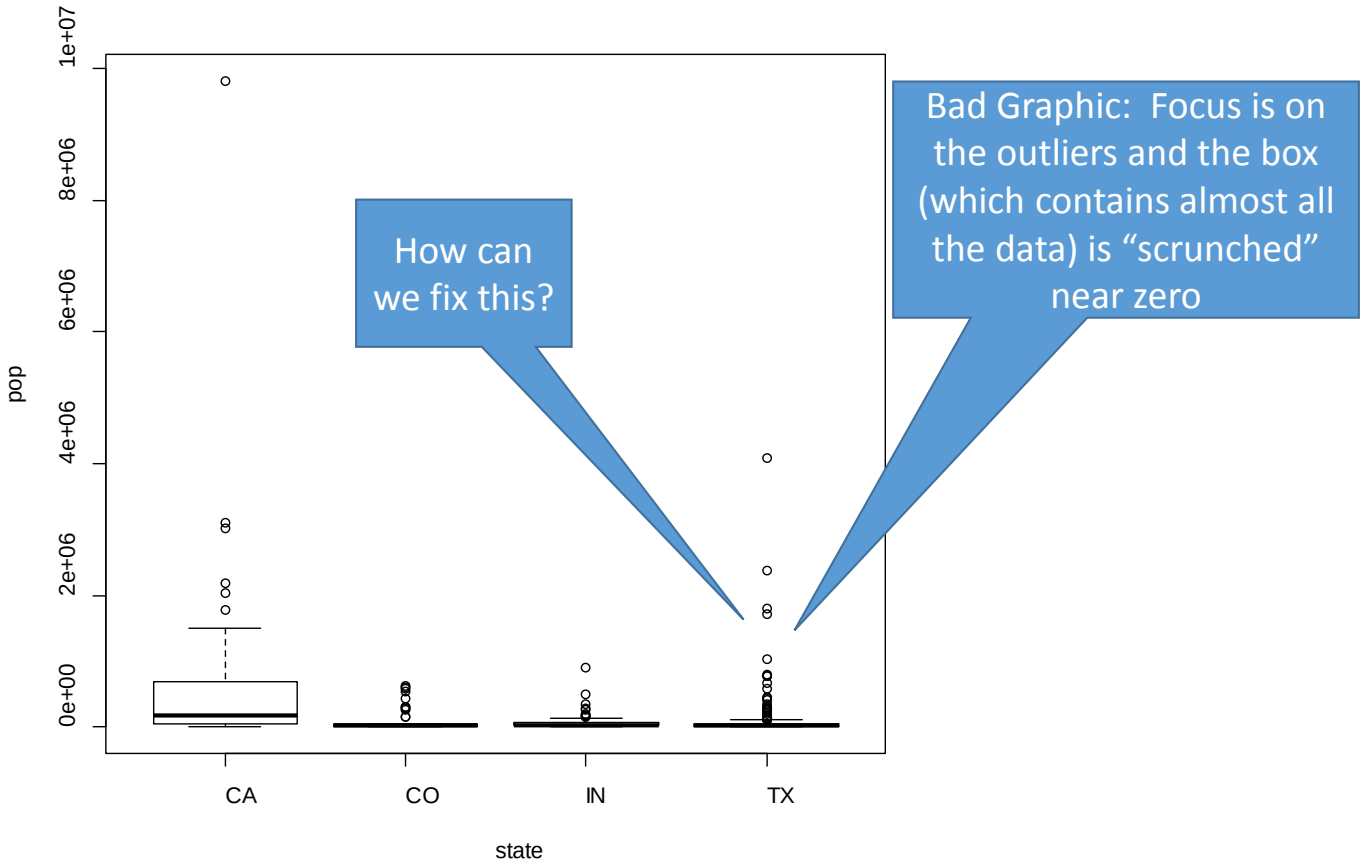


Compare batches using boxplots

- Data: total population of counties in 4 states in U.S.
- Questions: want to compare population distributions in 4 states
- Comparison: median, spread, symmetry, tail length, outliers
- R: *boxplot(formula)*
  formula: such as $y \sim grp$, where y is a numeric vector of data values to be split into groups according to the grouping variable *grp* (usually a factor).

Compare batches using boxplots

```
install.packages("noncensus")
library(noncensus)
data(counties)
?counties
head(counties)
b1=state=="CA"
b2=state=="CO"
b3=state=="IN"
b4=state=="TX"
b=b1|b2|b3|b4
dat=counties[b,]
plot(pop~state,data=dat) # Ooops
dat$state  # have to redo factor levels
dat$state=factor(as.character(dat$state))
plot(pop~state,data=dat)
```


pop
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
CA
CO
IN
TX
state
How can we fix this?
Bad Graphic: Focus is on the outliers and the box (which contains almost all the data) is "scrunched" near zero

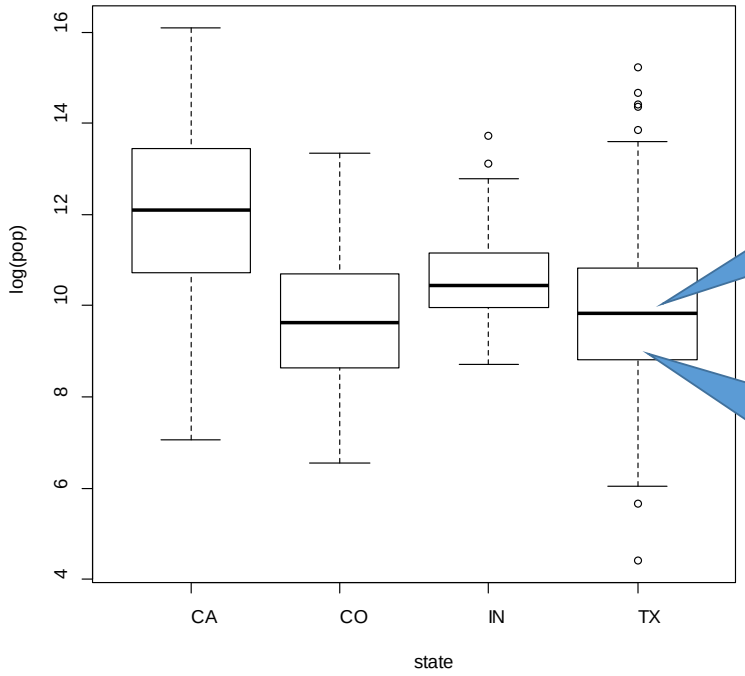
Compare batches using boxplots

- Data: total population of counties in 4 states in U.S.
- Questions: want to compare population distributions in 4 states
- Comparison: median, spread, symmetry, tail length, outliers
- R: *boxplot(formula)*
  formula: such as $y \sim grp$, where y is a numeric vector of data values to be split into groups according to the grouping variable *grp* (usually a factor).
- Dependency of spread on level, tendency for spread to increase as level does (belief: equal variability across batches)
- To promote equality: re-express or transform

Transformed Boxplots

Tukey's solution is to transform the data using a smooth monotonic increasing function like x^p or log(x)

```
plot(log(pop)~state,data=dat)
```



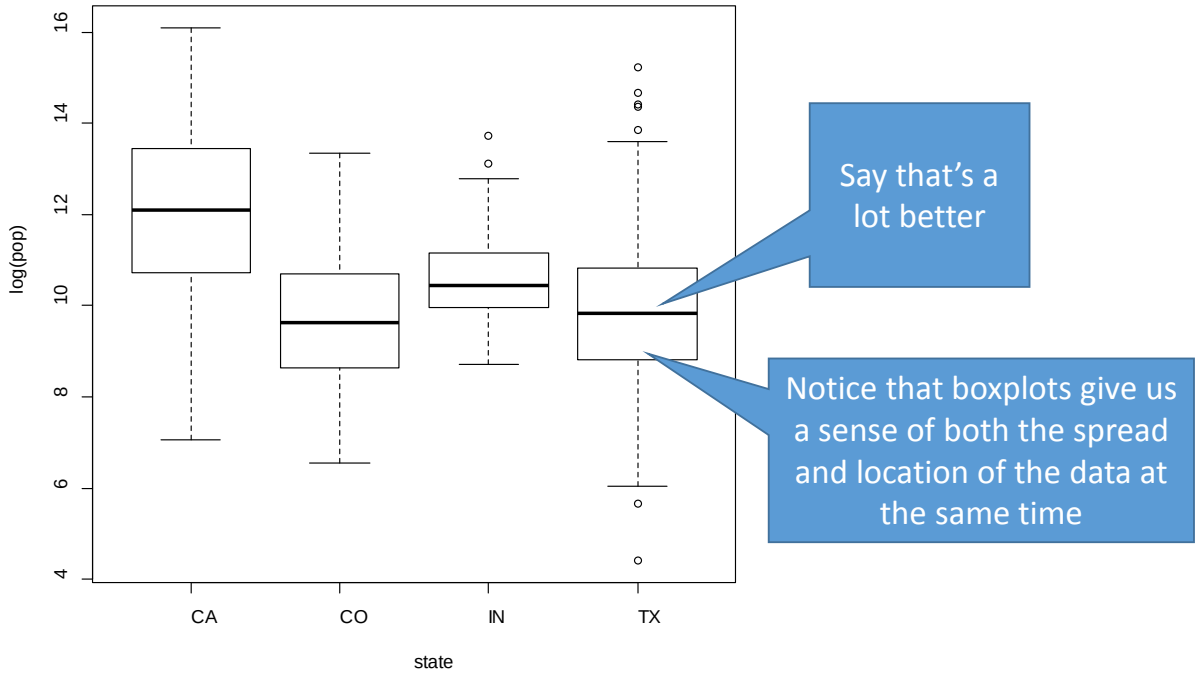
Say that's a lot better

Notice that boxplots give us a sense of both the spread and location of the data at the same time

Transformed Boxplots

Tukey's solution is to transform the data using a smooth monotonic increasing function like x^p or log(x)

plot(log(pop)~state,data=dat)

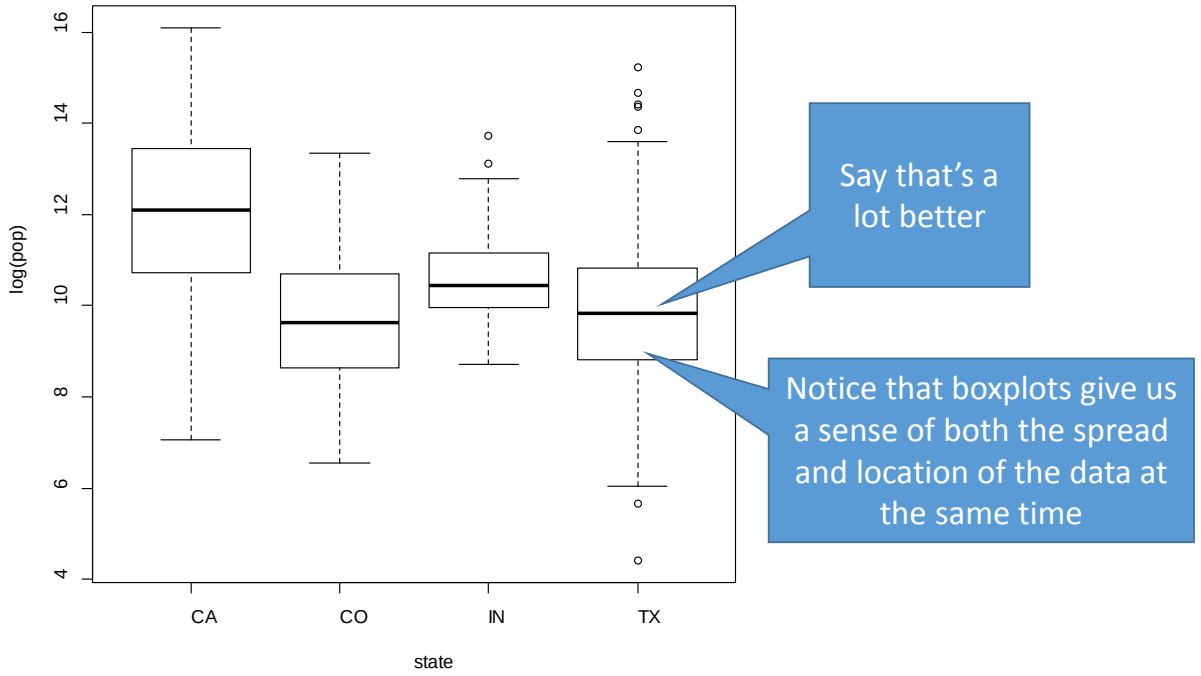


What about sample size?

Transformed Boxplots

Tukey's solution is to transform the data using a smooth monotonic increasing function like x^p or log(x)

plot(log(pop)~state,data=dat)



What about sample size? What about spread?

Notched boxplots

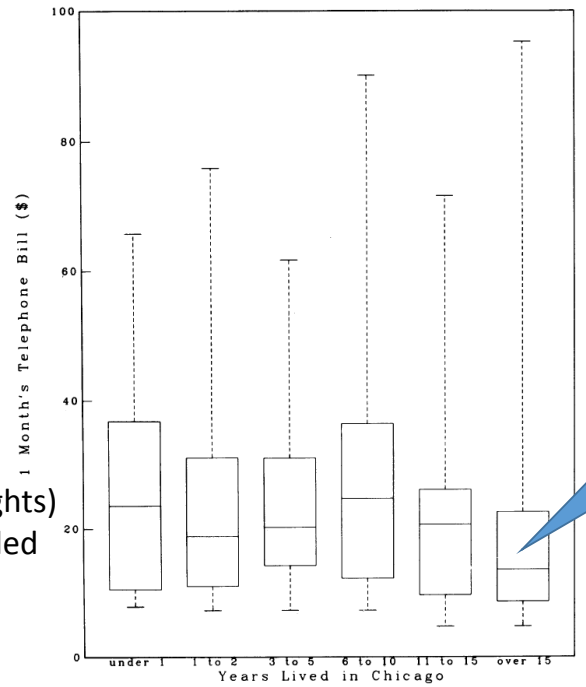
Reference: R McGill, JW Tukey, WA Larsen (1978): Variation of boxplots, The American Statistician, 32:12-16

- Boxplots can be misleading when sample size differ greatly
- Often want to compare locations (spread too)
- Enhance boxplots to better convey:
  (1) differences in sample sizes;
  (2) possible differences in location
  (2) possible differences in spread

Notched boxplots

Tukey discussed some of the problems with "ordinary Boxplots" using the Telephone Bill Data:

Years	*Customers*
less than 1	11
1 to 2	17
3 to 5	26
6 to 10	35
11 to 15	29
over 15	368



Is there less variation in final category or just a larger sample size?

Problems: 1. Users are led to believe that overall median is around \$20 (people assume equal weights)
2. Boxes have equal width. It would be better if we scaled the width to be proportional to sample size.

How can we improve the visual display to incorporate all the information?

Notched boxplots

Tukey Improvement #1: Scale the widths according to sample size

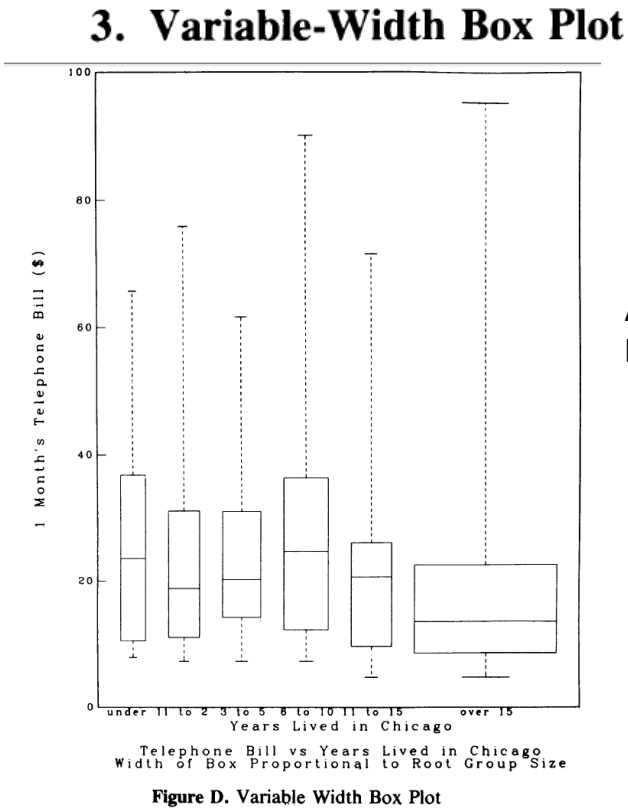


Figure D. Variable Width Box Plot

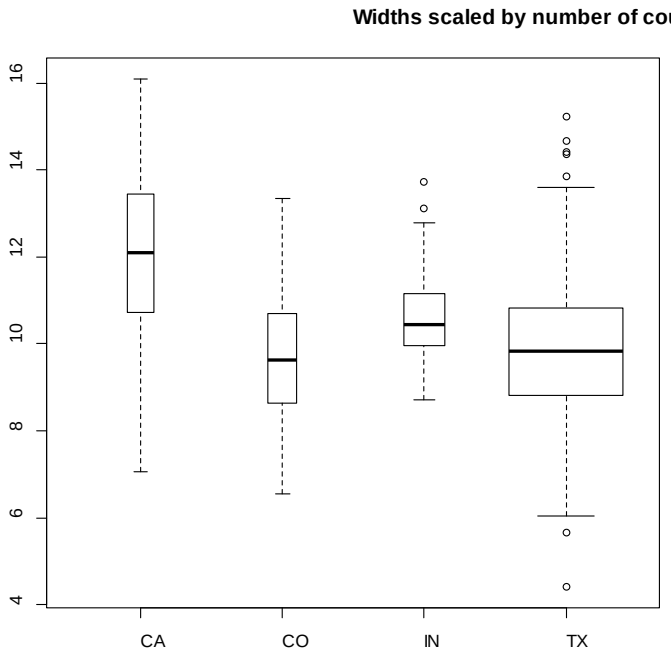
Accomplished in R with
boxplot(data, width = something)

Notched boxplots

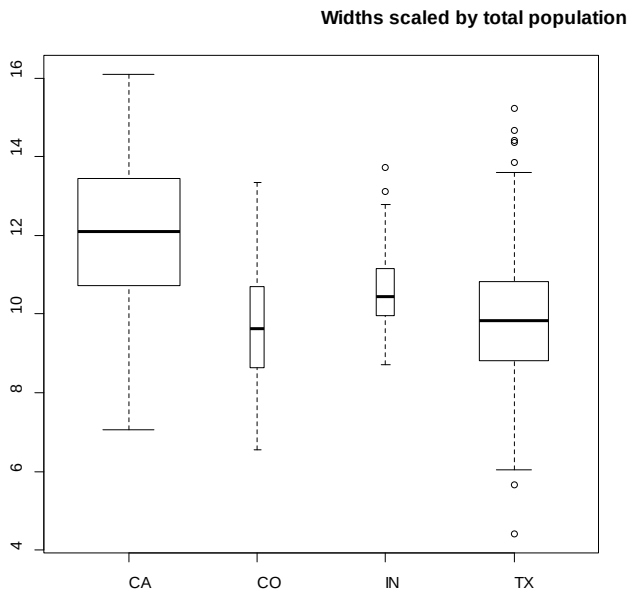
```
summary(dat$state)
 CA  CO  IN  TX
 58  64  92 254
```

```
> w=tapply(dat$pop,dat$state,sum)
> w
      CA       CO       IN       TX
37253956  5029196  6483802 25145561
```

```
boxplot(log(pop)~state,data=dat,width=summary(dat$state))
```




```
boxplot(log(pop)~state,data=dat,width=w)
```



Notched boxplots

Tukey Improvement #2: Put notches in the graph which allow the user to compare whether the medians of two different groups are statistically different.

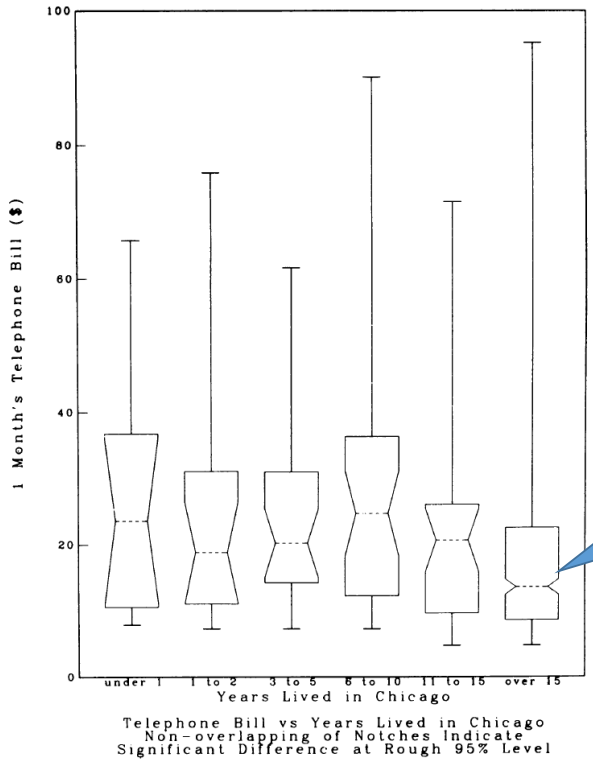


Figure E. Notched Box Plot

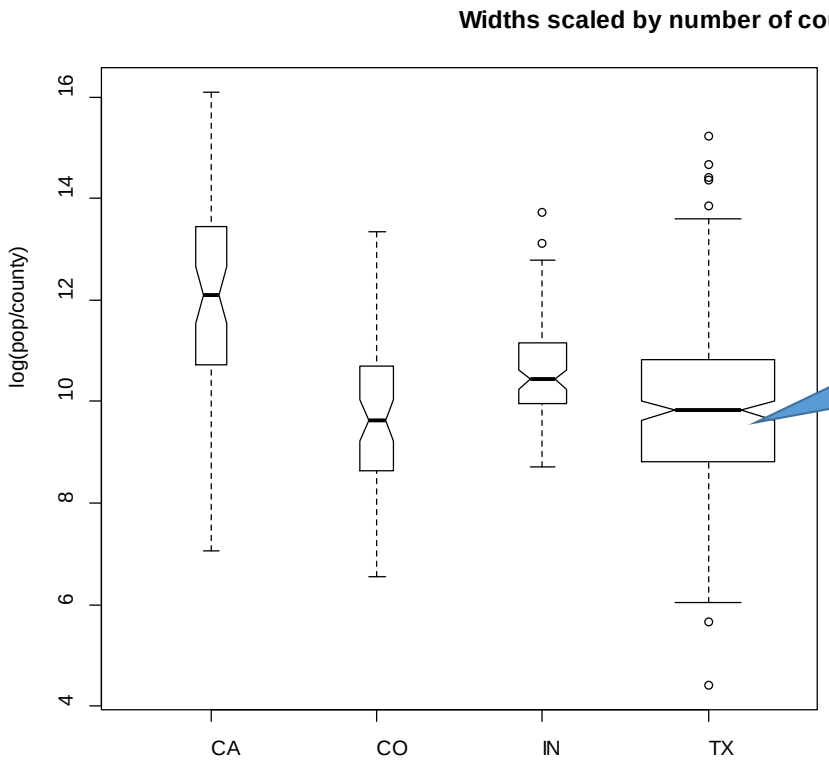
The medians for two different groups are statistically different at the 0.95% level if upper notch is lower than the lower notch of another group... or vice versa.

Constructing notched boxplots

- Same calculation as for boxplots
- Interval about median $(M - W, M + W)$: $W = 1.58(F_U - F_L)/\sqrt{n}$.
- Indent box corresponding to interval
- Scale box width proportional to $\sqrt{n}$
- Pair of nonoverlapping notches $\Rightarrow$ rough indication of significant difference between medians at 5% level
- R: *boxplot(data, notch=True, width)*

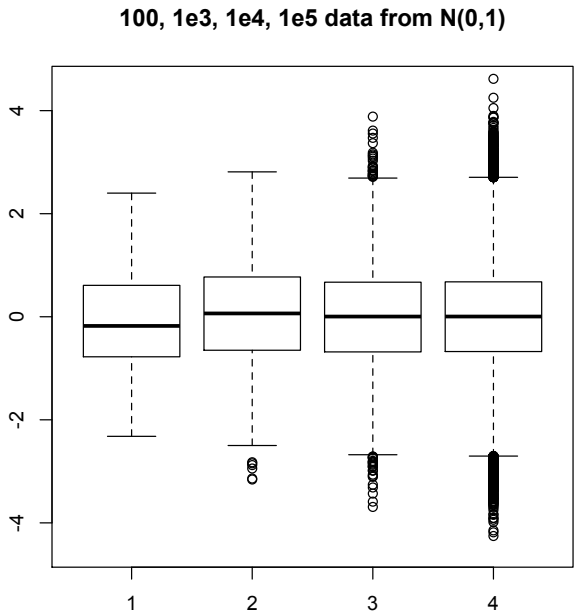
Constructing notched boxplots

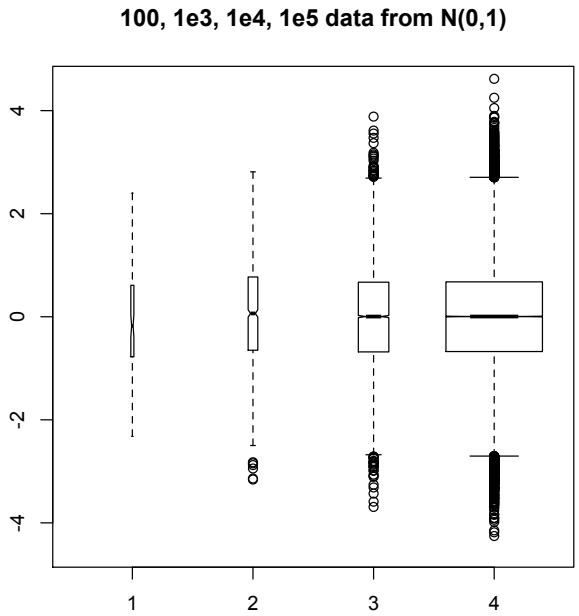
boxplot(log(pop)~state,data=dat,notch=TRUE,width=summary(
dat$state),main="Widths scaled by number of counties")



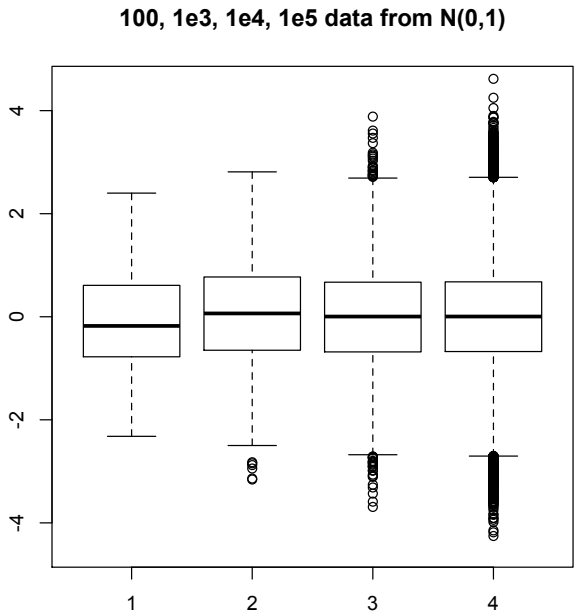
The width of notch incorporates information about Var(median)

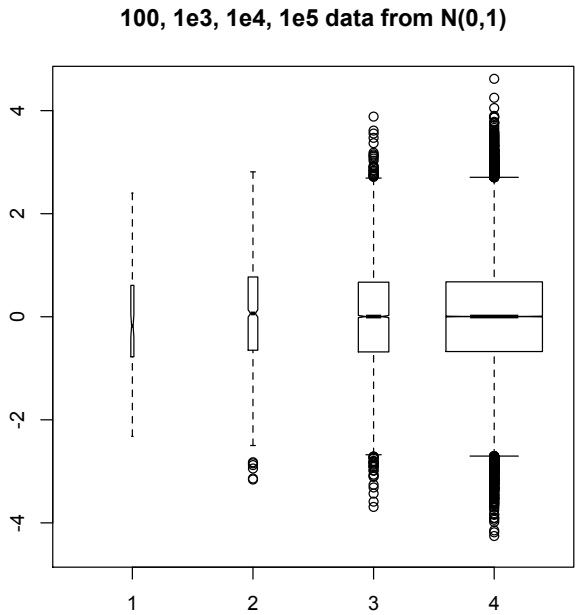
Example of Normal Boxplots





Example of Normal Boxplots





- In a regular boxplot, the only hint that the groups are different sizes is the number of outliers.
- The notched boxplots displays differences in group size with variable box width.
- The notched boxplots also displays an inferentially meaningful quantity: the error associated with the estimate of the median.

$w = 1.58IQR/\sqrt{n}$, where does 1.58 come from?

1. Gaussian data: $\text{SD}(\text{median}) = \text{SD}(M) = \sqrt{\pi/2} \cdot \sigma/\sqrt{n}$
2. $\text{E}(\text{F-spread}) = (0.6745\sigma - (-0.6745\sigma)) \approx 1.35\sigma$
3. $\text{SD}(M) \approx 1.25 \times (IQR/1.35) \times 1/\sqrt{n}$
4. For 95% CI around each median: $M \pm C \cdot SD(M)$, $C \approx 1.96$
5. We want rough 95% CI for difference, $M_1 - M_2$:

$$SD(M_1 - M_2) \approx [SD^2(M_1) + SD^2(M_2)]^{1/2}$$

6. If $n_1 = n_2$,

$$SD(M_1 - M_2) \approx \sqrt{2}SD(M_1) = 1.414 \times SD(M_1)$$

and 95% CI for difference is

$$(M_1 - M_2) \pm 1.96 \times \sqrt{2} \times SD(M_1)$$

$w = 1.58IQR/\sqrt{n}$, where does 1.58 come from?

7. For notches around each median, we could use $1/2$ this value

$$\text{notch length} \approx 1.96 \times \sqrt{2} \times SD(M_1) \times 1/2$$

$$\Rightarrow C = 1.96 \times \sqrt{2} \times 1/2 = 1.386$$

8. If $n_2 = n_1/3$ (one batch is 3 times larger than another)

$$SD(M_1 - M_2) \approx \sqrt{1+3} \times SD(M_1) = 2 \times SD(M_1)$$

$$\text{notch length} \approx 1.96 \times \sqrt{1+3} \times SD(M_1) \times 1/2$$

$$\Rightarrow C = 1.96 \times 2 \times 1/2 = 1.96$$

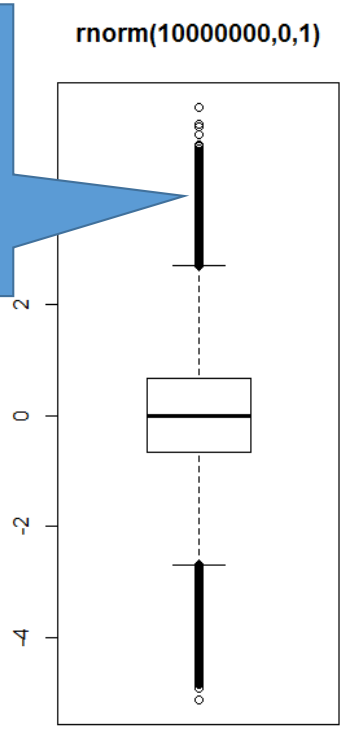
9. 1.7 was chose as a compromise between the cases where $n_1 \approx n_2$ and $n_1 \approx 3n_2$.
10. Putting all together, $M \pm W$ where

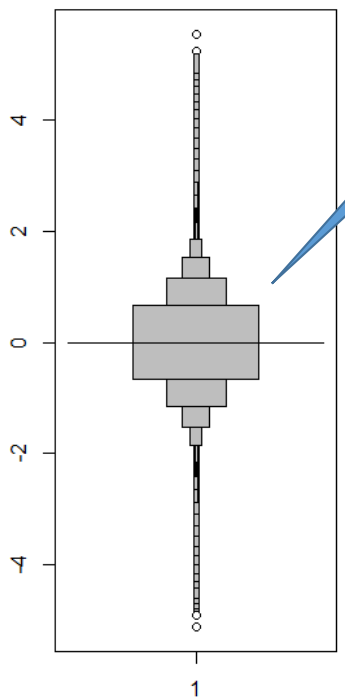
$$W = 1.7 \times 1.25 \times IQR/1.35 \times 1/\sqrt{n} \times 1/2 = 1.58IQR/\sqrt{n}$$

Boxplot and Letter Value Plot side by side

```
x=rnorm(10000000,0,1)
par(mfrow=c(1,2))
boxplot(x,main="rnorm(10000000,0,1)")
lvplot(x)
```

Even if the data come from a nice symmetric distribution, a lot of data will have a lot of outliers.





Much nicer

Upper sixteenth
Upper eighth
Upper fourth
M
Lower fourth
Lower eighth
Lower sixteenth

Graphical Displays for large datasets

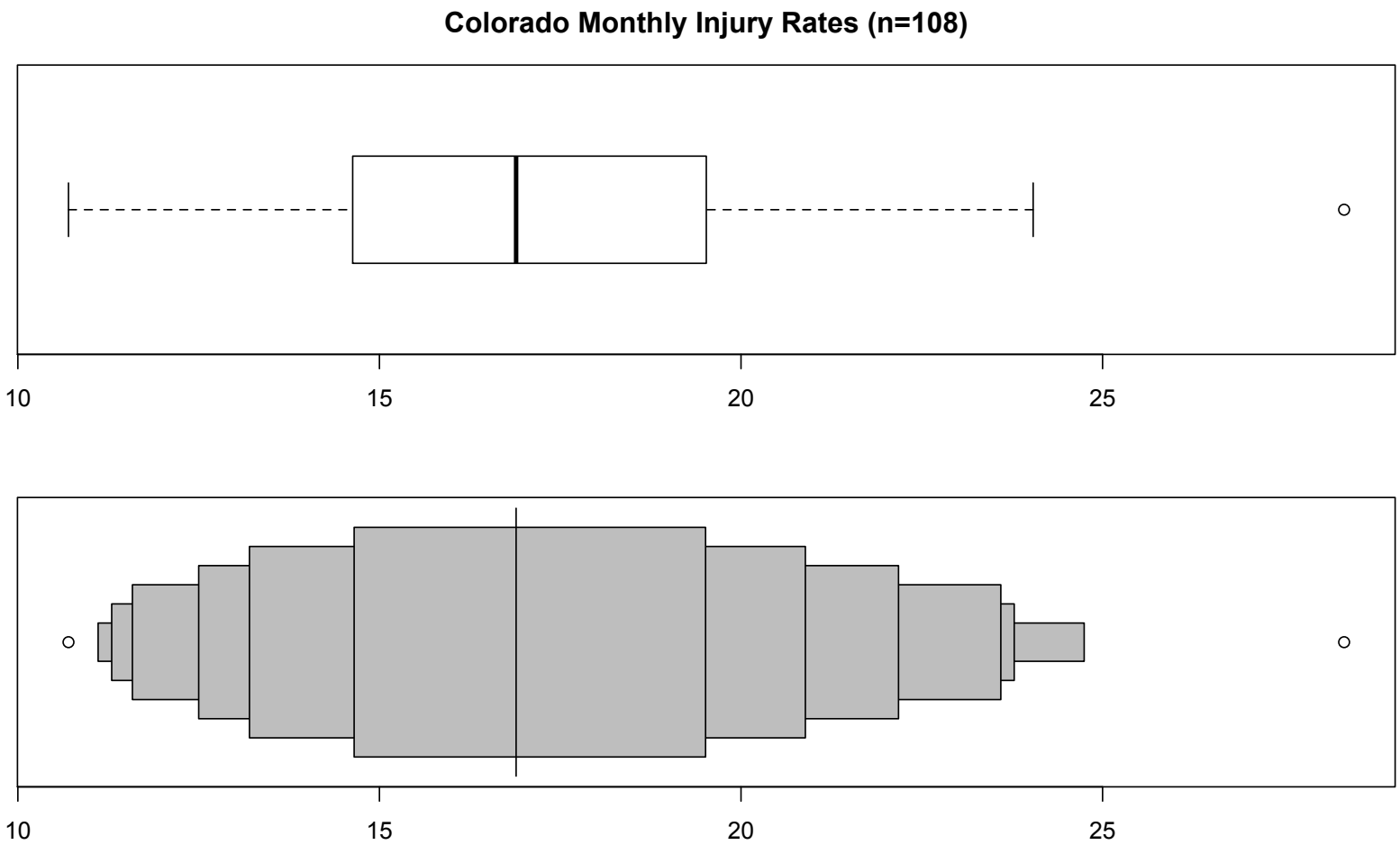
Great value in displaying large data sets

- Quick summary of data characteristics without being overwhelmed with too much detail
- Approximate location, spread, shape of distribution
- Observe outliers
- Association among variables (especially in pairs)
- "The greatest value of a picture is when it forces us to notice what we never expected to see" (Tukey 1977)

Letter value box plots

- Small datasets: Limited information about tails
- Boxplots show fourths, *extent* of data beyond fourths
- Large dataset: Tail quantiles more reliable
- $\Rightarrow$ Extend box plots to include more letter values beyond median, fourths
- "Stopping Rules": How many LVs to show?
- How to display letter values?
- Which observations are labeled as "outliers"?
- **Plot still shows only actual data values**
- Still unable to show multi-modality.

Example

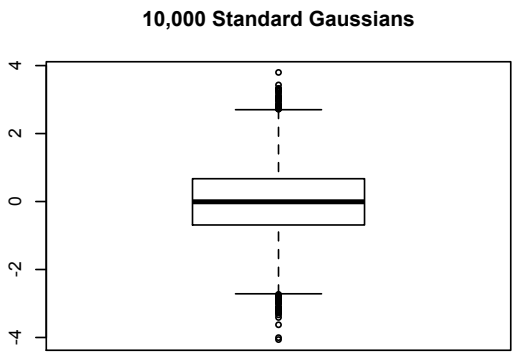


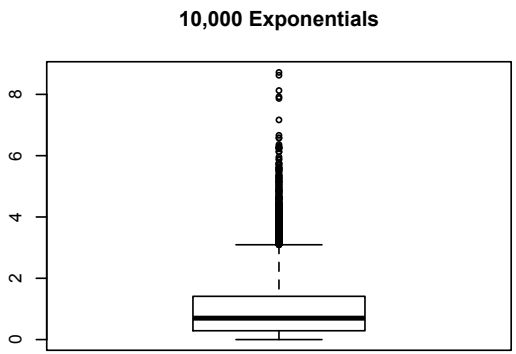
Stopping rules and labeled outliers

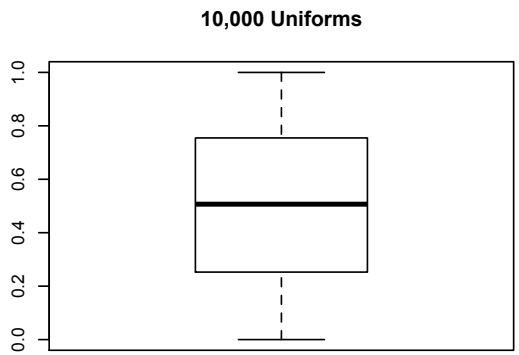
1. Fix # of labeled outliers: $k = \lfloor \log_2 n \rfloor - 4$
2. Fix percentage of data (e.g., 0.5-1.0%) as labeled outliers
3. "Trustworthiness" of LV_i as estimate of corresponding quantile: Stop at k if uncertainty for LV_k extends beyond or into LV_{k-1} (e.g. '95% limits' on LV_i crosses over LV_{i-1}): $k = \lceil \log_2 n - \log_2(4z^2_{1-\alpha/2}) \rceil + 1$

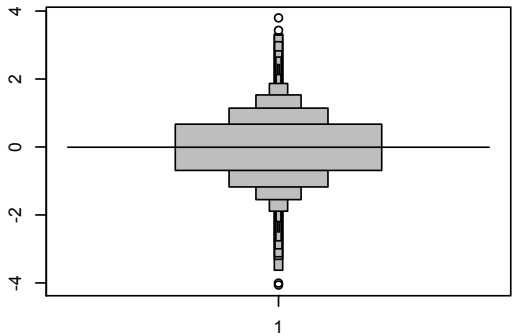
All three rules lead to roughly similar answers.

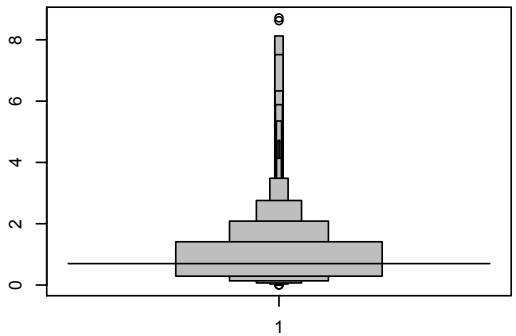
Examples: Gaussian, Exponential, Uniform ($n = 10,000$)

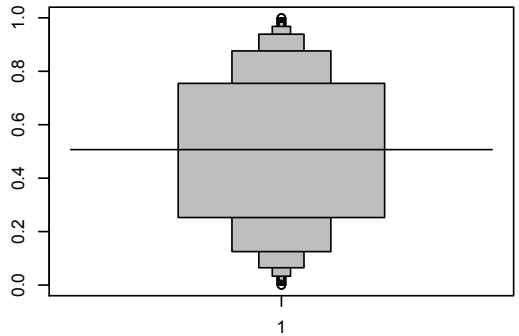




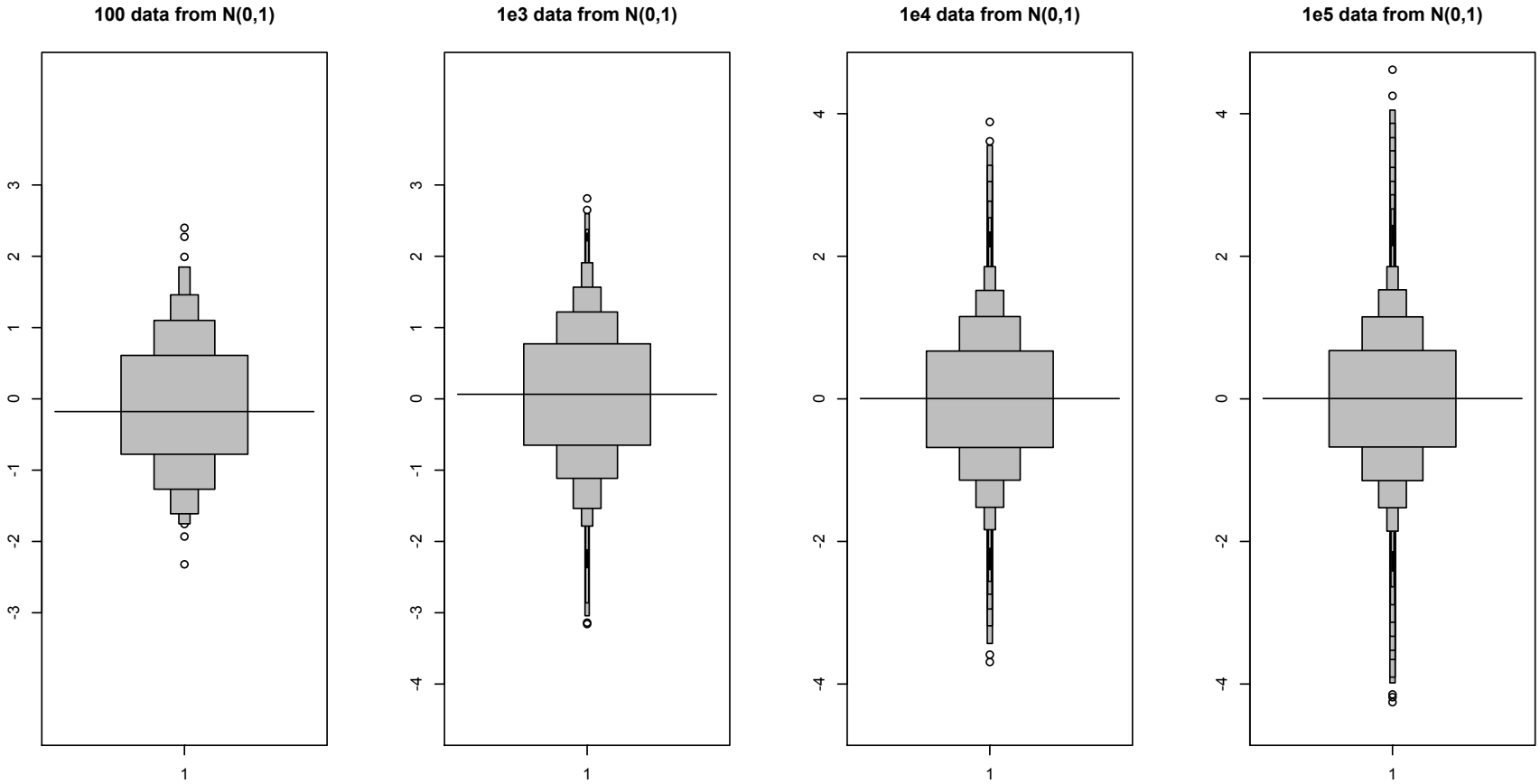








Examples: Gaussian ($n = 100, 1000, 1e4, 1e5$)



R funtion: *lvplot(data)*

Some R Exercises

```
#source KK's lval programs
source("lvalprogs.r")

library(MASS)

#Read in data file warpbreaks
attach(warpbreaks)
x <- breaks

# Letter Values
lval(breaks)

# Boxplot vs Stackbox
par(mfcol=c(2,2))
boxplot(breaks,horizontal=T,main="Number of Breaks in yarn during weaving")
stackbox(breaks)
boxplot(log(breaks),horizontal=T,main="Log Breaks")
stackbox(log(breaks))

# Boxplot vs Lvplot
par(mfrow=c(2,2))
boxplot(breaks,horizontal=F,main="Number of Breaks in yarn during weaving")
lvplot(breaks); title("Letter value plot")
boxplot(log(breaks),horizontal=F,main="Log Breaks")
lvplot(log(breaks + 1)); title("Letter value plot")
```

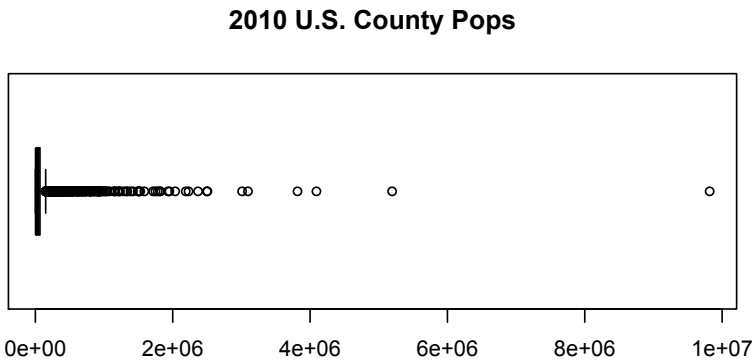
Some R Exercises

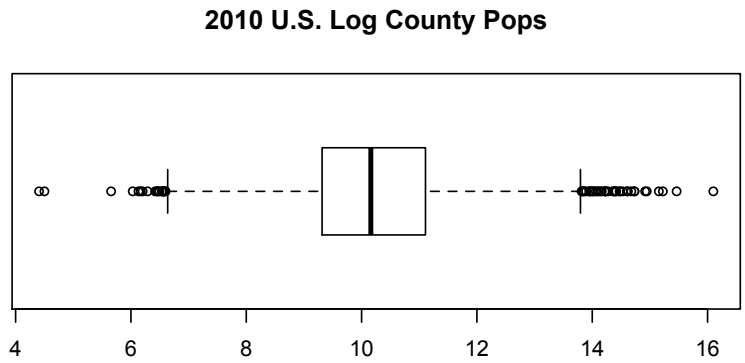
```
# QQ lines
par(mfrow=c(2,2))
qqnorm(breaks,ylab="Breaks",main="(A): Breaks"); qqline(breaks)
qqnorm(log(1+breaks),ylab="Log(1+Breaks)",
       main="(B): Logarithm"); qqline(log(1+breaks))
qqnorm(sqrt(breaks),ylab="sqrt(Breaks)",
       main="(C): Square root"); qqline(sqrt(breaks))
qqnorm(breaks^.25,ylab="Breaks^.25",
       main="(D): Fourth roots"); qqline(breaks^.25)

# LV QQ lines
par(mfrow=c(2,2))
qqnormLV(breaks,main="(A): Breaks (LVs)")
qqnormLV(log(1+breaks), main="(B): Logarithm (LVs)")
qqnormLV(sqrt(breaks), main="(C): Square root (LVs)")
qqnormLV(breaks^.25, main="(D): Fourth roots (LVs)")
#dev.off()
```

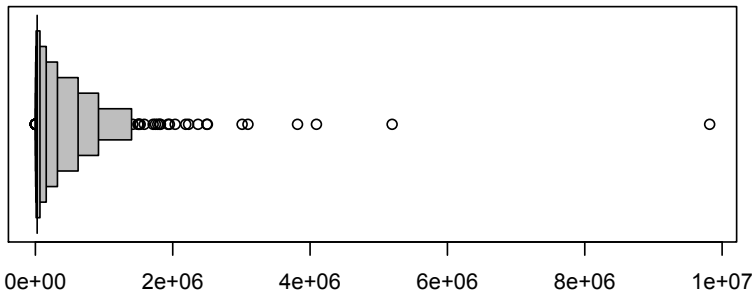
Illustrations of large dataset display

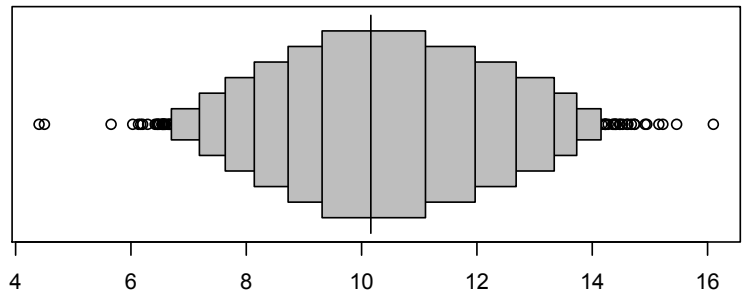
1. Population of 3143 U.S. counties (combined)
2. Population of 290 counties in seven states
3. Association between message size variables collected on 135,605 Internet sessions (KK + EJW)

Illustrations of large dataset display

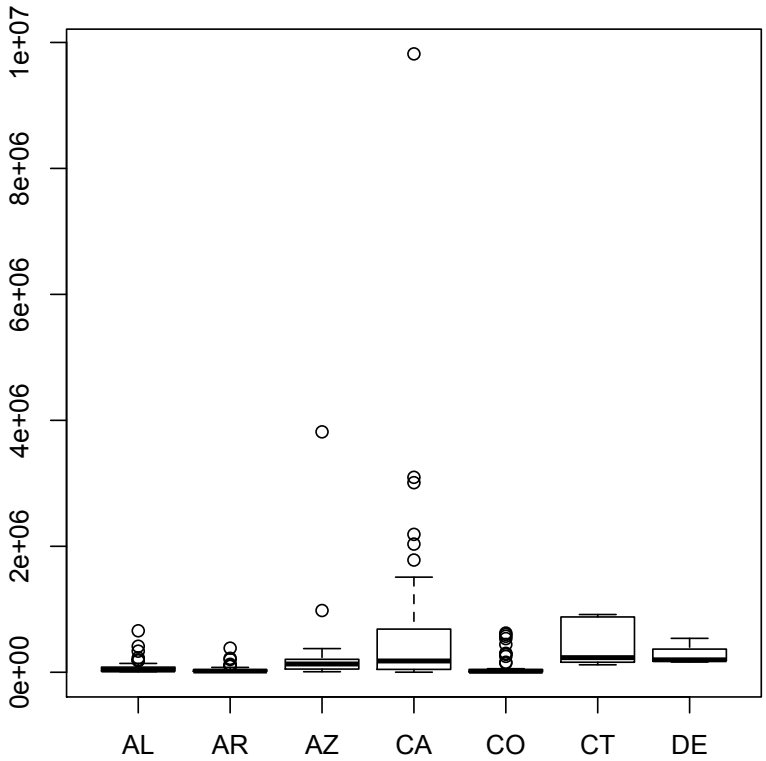


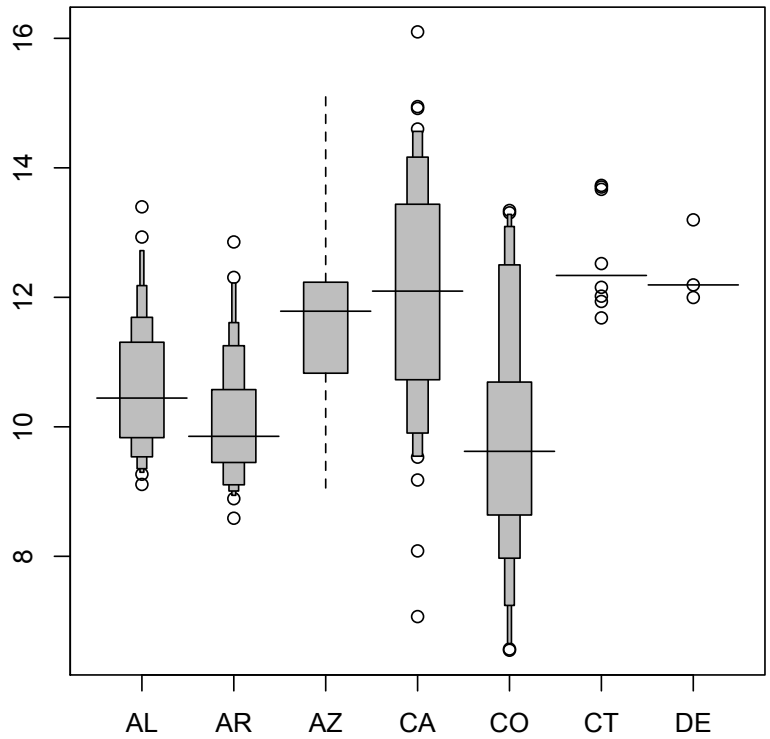






Illustrations of large dataset display





Illustrations of large dataset display

Internet sessions data

Measures of duration and size of 135,605 Internet sessions, collected during 1 hour of traffic

Duration: Length of message (in seconds)

Size: Number of bytes, number of packets

Goal: Detect "Unusual" activity

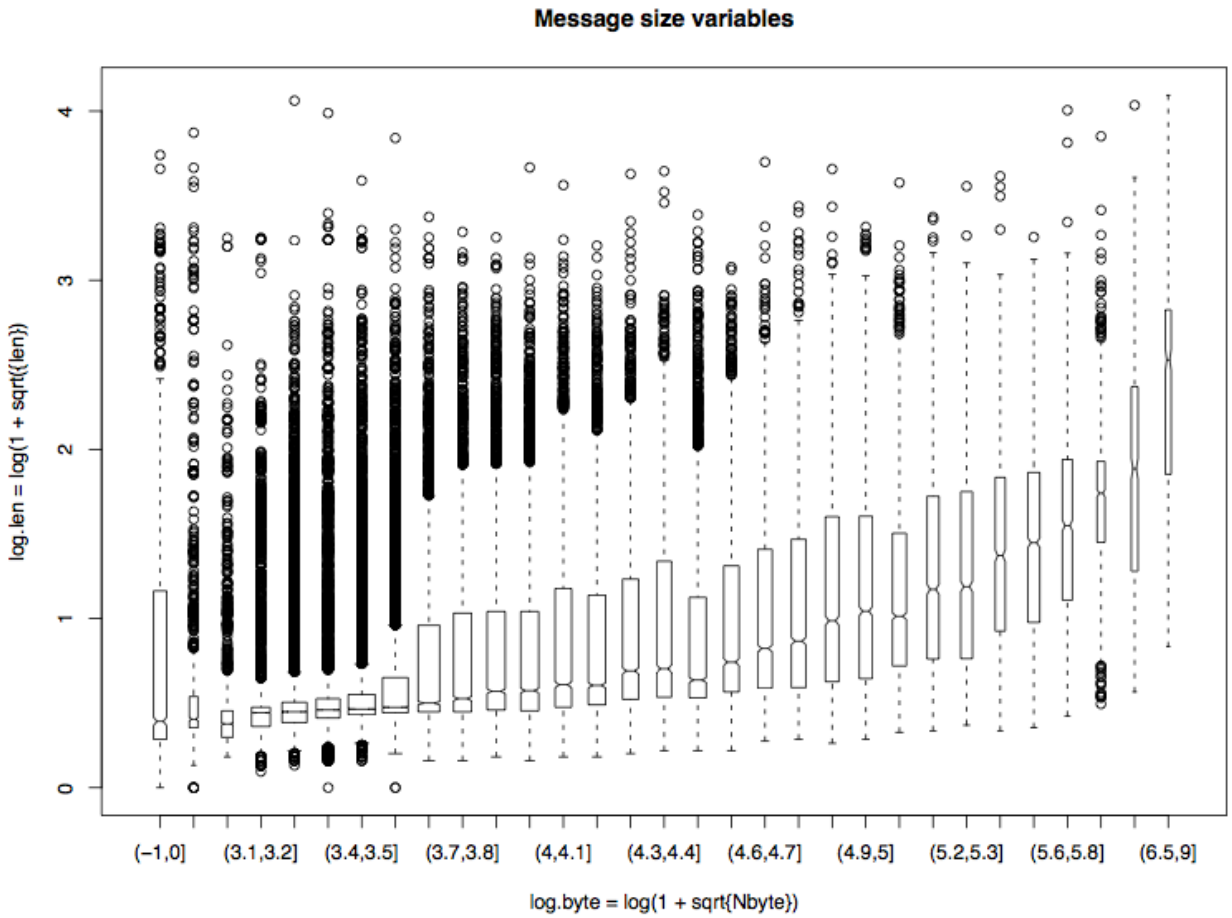
EDA reveals:

- Nonlinear associations between duration and size variables
- Very skewed, long-tailed distributions; need to transform

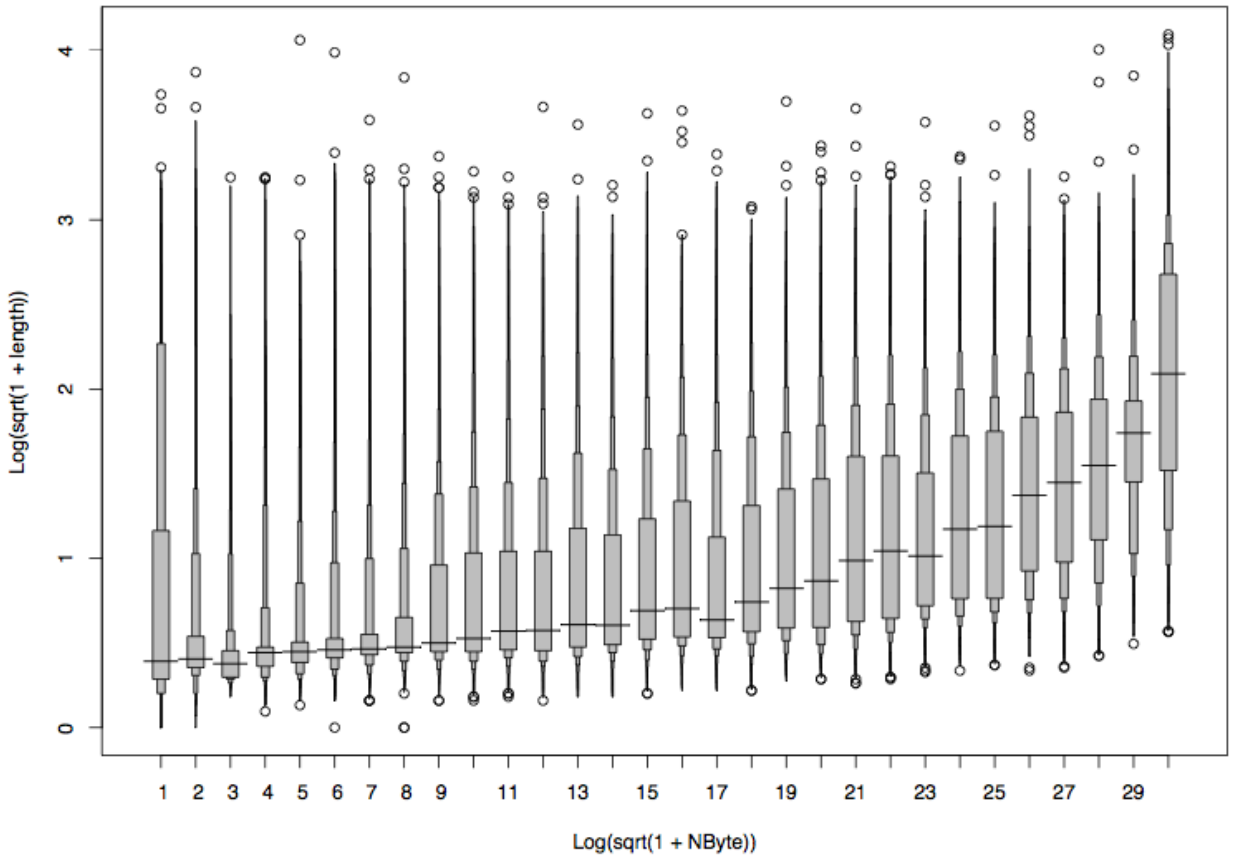
$$\texttt{log.byte} = \log(1 + \sqrt{\text{No. bytes}})$$
$$\texttt{log.len} = \log(1 + \sqrt{\text{duration}})$$

(KK + EJW, *CSDA* 2006)

Illustrations of large dataset display



Illustrations of large dataset display



Bivariate extensions of box plot

Challenge: bivariate analog of letter values

- *ldepth* (location depth) of point $\theta \in \mathcal{R}^p$ = smallest number of observations contained in any closed halfspace (half plane when $p = 2$) with boundary line through θ.
- When $p = 2$, "draw" (conceptually) all possible lines through θ
- For each "line", count # points on either side
- $ldepth(\theta)$ = smallest of these numbers
- Halfspace median $\equiv$ value θ_M for which this number is the largest (or center of gravity of all θ's that satisfy criterion)

Bivariate extensions of box plot

Bagplot

- A 2D Boxplot Extension
- The *bagplot*(x, y) function in the "aplpack" package.
- The bag contains 50% of all points.
- The bivariate median is approximated.
- The fence separates points in the fence from points outside. Outliers are displayed.

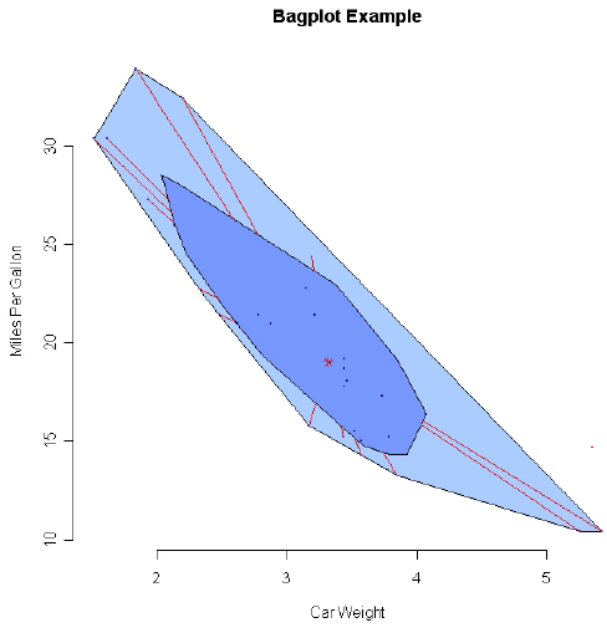
Example: `data(mtcars)` in `R`

1974 *Motor Trend* data on 32 1973-74 cars

1.	`mpg`	6.	`gear` (#forward gears)
2.	`cyl`	7.	`qsec` ($\frac{1}{4}$-mile time)
3.	`disp`	8.	`drat` (rear axel ratio)
4.	`hp`	9.	`vs` (V/S)
5.	`wt`	10.	`carb` (#carburettors)
11.	`am` Transmission (0=auto, 1=manual)		

Bivariate extensions of box plot

```
library(aplpack)
attach(mtcars)
bagplot(wt,mpg, xlab="Car Weight", ylab="Miles Per Gallon",
main="Bagplot Example")
```



Summary

Letter value box plots:

- are especially appropriate for large data sets
- are based on actual data values
- are simple to compute from letter value displays
- impart detail buried in "whiskers" of conventional boxplots
- show fewer outliers than do box plots ($\approx$0.7%)
- do not depend on smoothing parameter (density estimate)

lvplot: R code

```
lvplot <- function(x,tag,ylabname="", ...) {
# Karen Kafadar, 2006
# this may be slow for many groups
  if(missing(tag)) tag <- rep(1,length(x))
  y <- x[!is.na(x)]
  tag <- tag[!is.na(x)]
  groupid <- unique(tag)
  ngroup <- length(groupid)
  ni <- table(tag)
  xlabname <- as.character(groupid)
   plot(c(0.5,ngroup+0.5),range(y,na.rm=T),xlab="",axes=F,
     type="n", ylab=ylabname,...)
  box()
  axis(1,at=1:ngroup,labels=xlabname)
  axis(2,at=pretty(y))
```

lvplot: R code

```
  for (j in 1:ngroup) {
    x <- y[tag==groupid[j]]
    n <- ni[j]
# ensure that k is at least 2, so at least fourths are shown
    k <-  1+max(2, 1 + ceiling(log(n,2) - log(4*1.96^2,2)))
    lval.x <- lval.sub(x)
    qu <- c(rev(lval.x[1:k,2]),lval.x[1:k,3])
    med <- qu[k]
    lfence <- 4*lval.x[2,2] - 3*lval.x[2,3]
    ufence <- 4*lval.x[2,3] - 3*lval.x[2,2]
    lfourth <- lval.x[2,2]
    ufourth <- lval.x[2,3]
    lower.adj <- ifelse(min(x) < lfence,
    min(x[x > lfence]), min(x))
    upper.adj <- ifelse(max(x) > ufence,
    max(x[x < ufence]), max(x))
```

lvplot: R code

```
\tiny{
    # draw boxes:
    wid <- 1/2^(k:1) - 0.01
    if (n < 30) {   # i.e., k=2: draw boxplot
      if (n < 10) {
         points(rep(j,n),x)
         segments(j-.49, med, j+.49, med)  # line for median
      }
      else {
         rect(j-wid[k-1], lfourth, j+wid[k-1], ufourth,
         col="grey")
         segments(j-.49, med, j+.49, med)  # line for median
         out <- (x < lower.adj) | (x > upper.adj)
         points(rep(j,sum(out)),x[out])
         segments(j,lfourth,j,lower.adj,lty=2)
         segments(j,ufourth,j,upper.adj,lty=2)
     }
    }
```

lvplot: R code

```
    else {
      # out <- (x < min(lfence, qu[1])) |
      (x > max(ufence, qu[2*k]))
      out <- (x < qu[1]) | (x > qu[2*k])
      points(rep(j,sum(out)),x[out])
      for (i in 1:k)
      rect(j-wid[i], qu[i], j+wid[i], qu[2*k-i+1], col="grey")
    }
  }
}
```

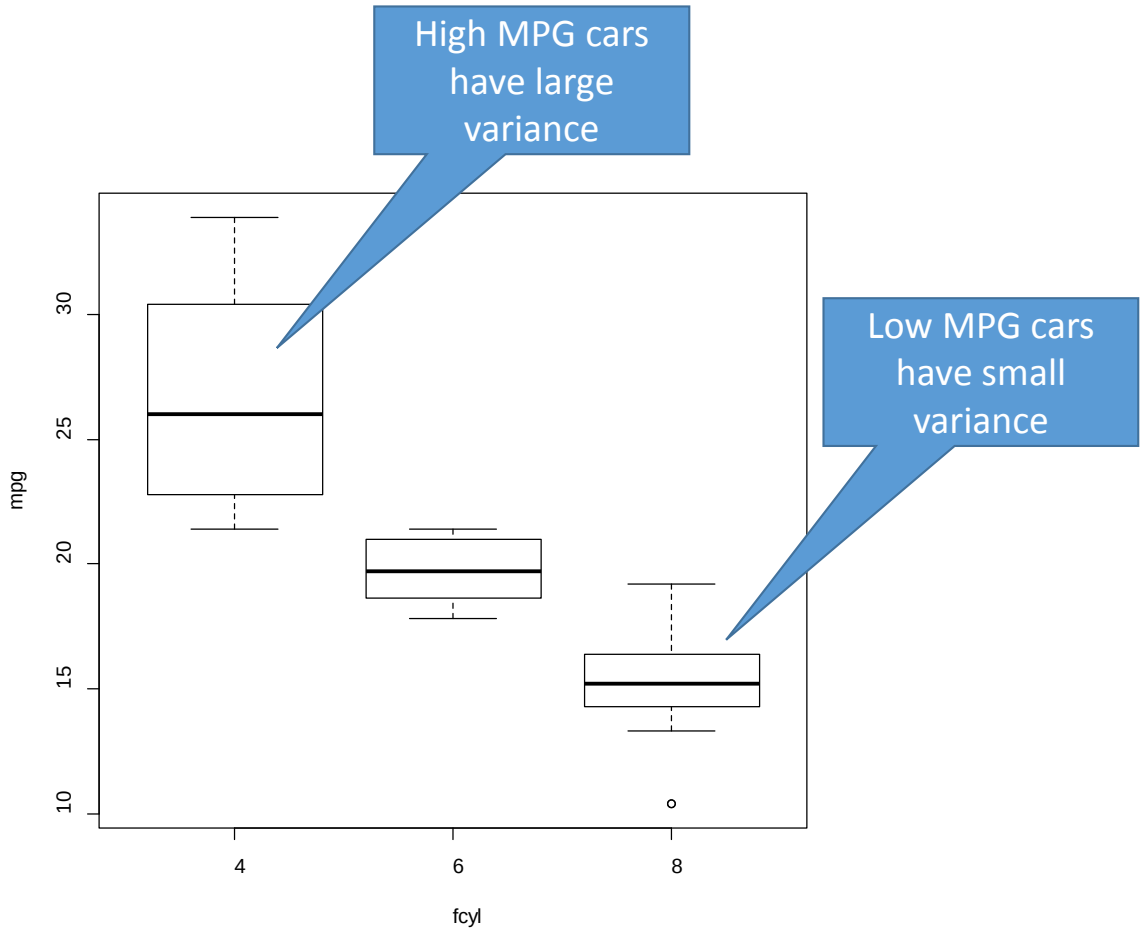
Transforming Data: Spread vs Level Plot

Data Transforms

- Tukey drew box plots a lot
- As he did this for a lot of problems he noticed that as the median increased the spread of the data also tended to increase
- Examples he used in text book were:
  - Box plots of 10 largest cities in countries
  - Box plots of MPG in cars (large variablility with high mpg cars)

Example

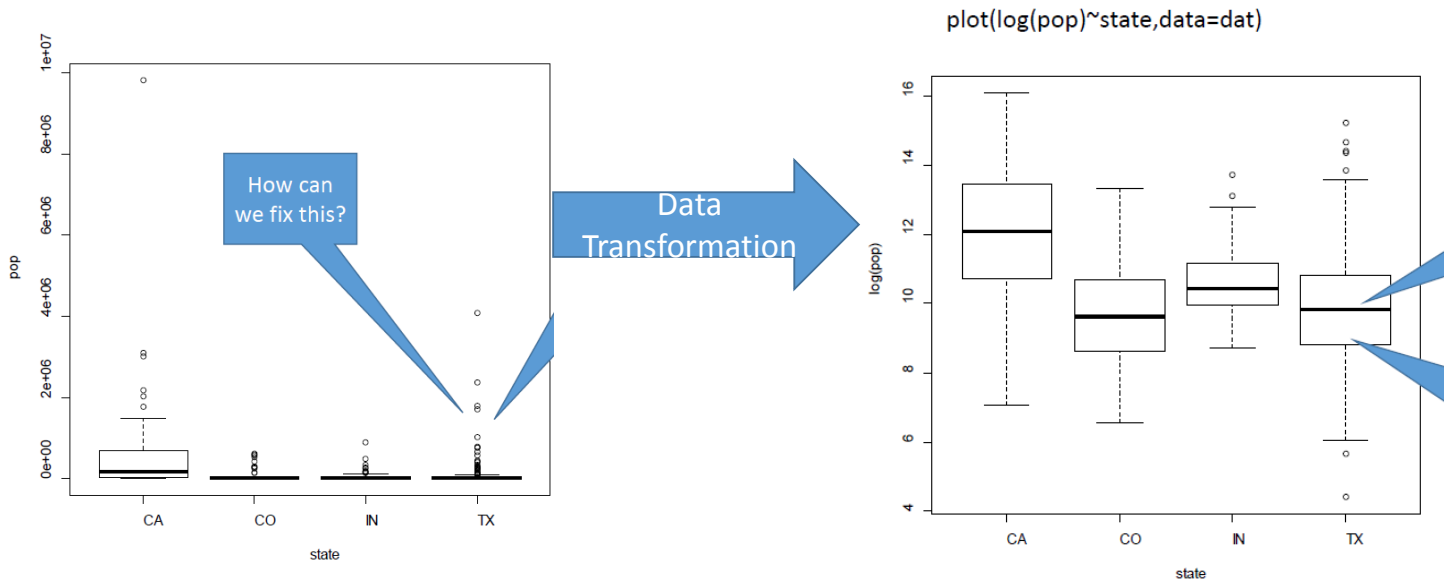
```
mtcars
fcyl=factor(mtcars$cyl)
mpg=mtcars$mpg
plot(mpg~fcyl)
```



Reasons for Transforms

Reasons for Transforming Data:

1. Transform data to enhance interpretability.
2. Transform data so that it is more symmetric
3. Transform data so that we have a uniform spread
4. Transform data for straightness (linearity)



Re-expression or transformation

- When?
  Batches comparison show a systematic relationship between spread and level (want to eliminate dependency for better visual exploration or future analysis, e.g., ANOVA
- What?
  replace x by x^p, $p = 0$, use $\log x$ (Explanation in Ch4)
- How?
  Spread-versus-level plot $\rightarrow$ appropriate power transformation (fuller discussion in Ch4 & Ch8)

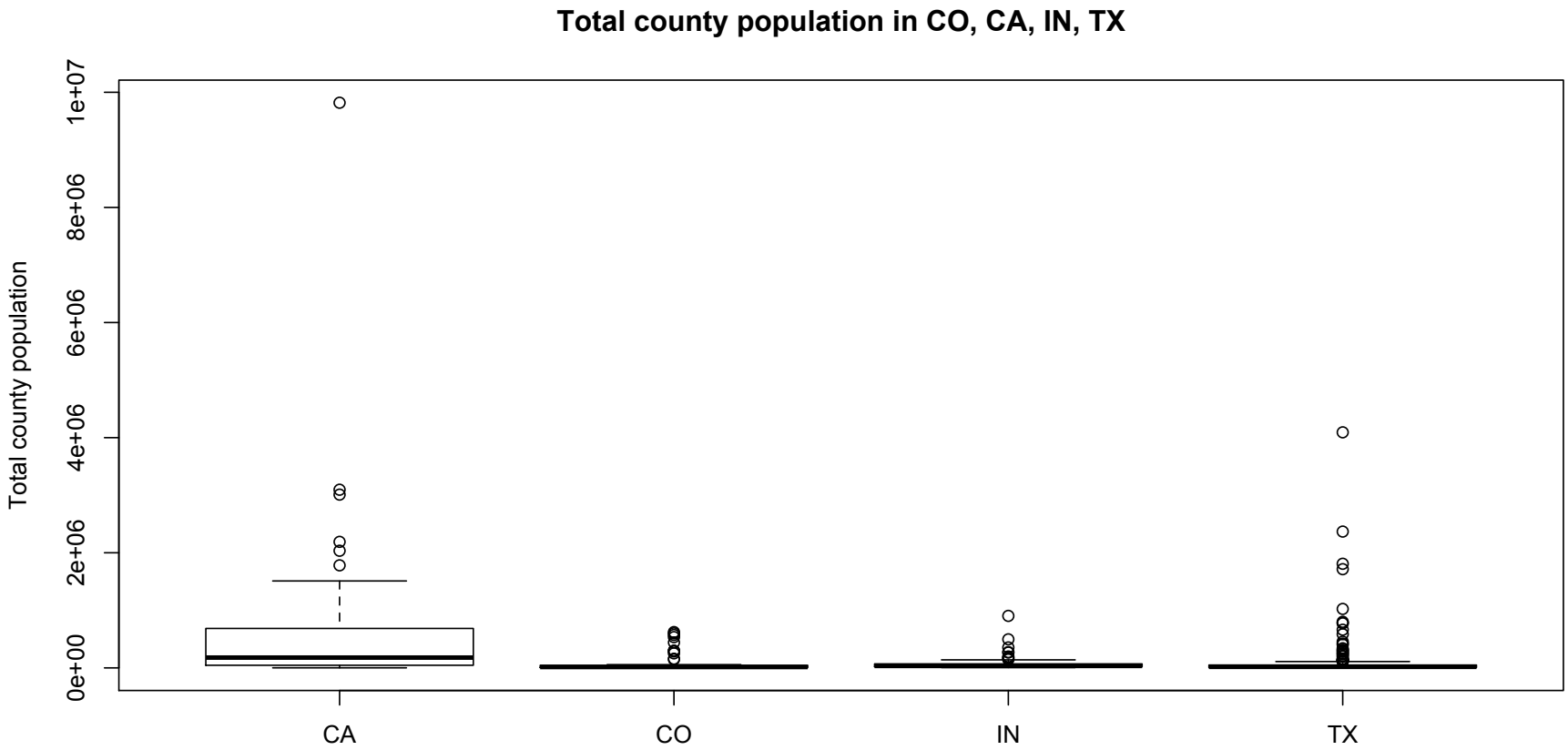
Construction of spread-versus-level plots

- Understand relationship between spread and level
  Suppose $d_F = CM^b$, $\log d_F = \log C + b \log M = k + b \log M$
- Spread-versus-level plot: a plot of $\log d_F$ against $\log M$ for all batches
- b is slope, $p = 1 - b$ is the approximate value of exponent for a power transformation of x to stabilize spread. $b = 1$, $p = 0$, use logarithms. $b = 1/2$, $p = 1/2$, square root transformations. (simplicity, interpretability)
- Recheck w/ (d_F vs M) plot
- Make choices: (1) equalize spread; (2) subject-matter explanation.
  demography (log $\rightarrow$ linear): (advantages of linear growth: interpretability, ease of detecting departure from fit, convenience in interpolation); (3) choose a power that's an integer multiple of 1/2

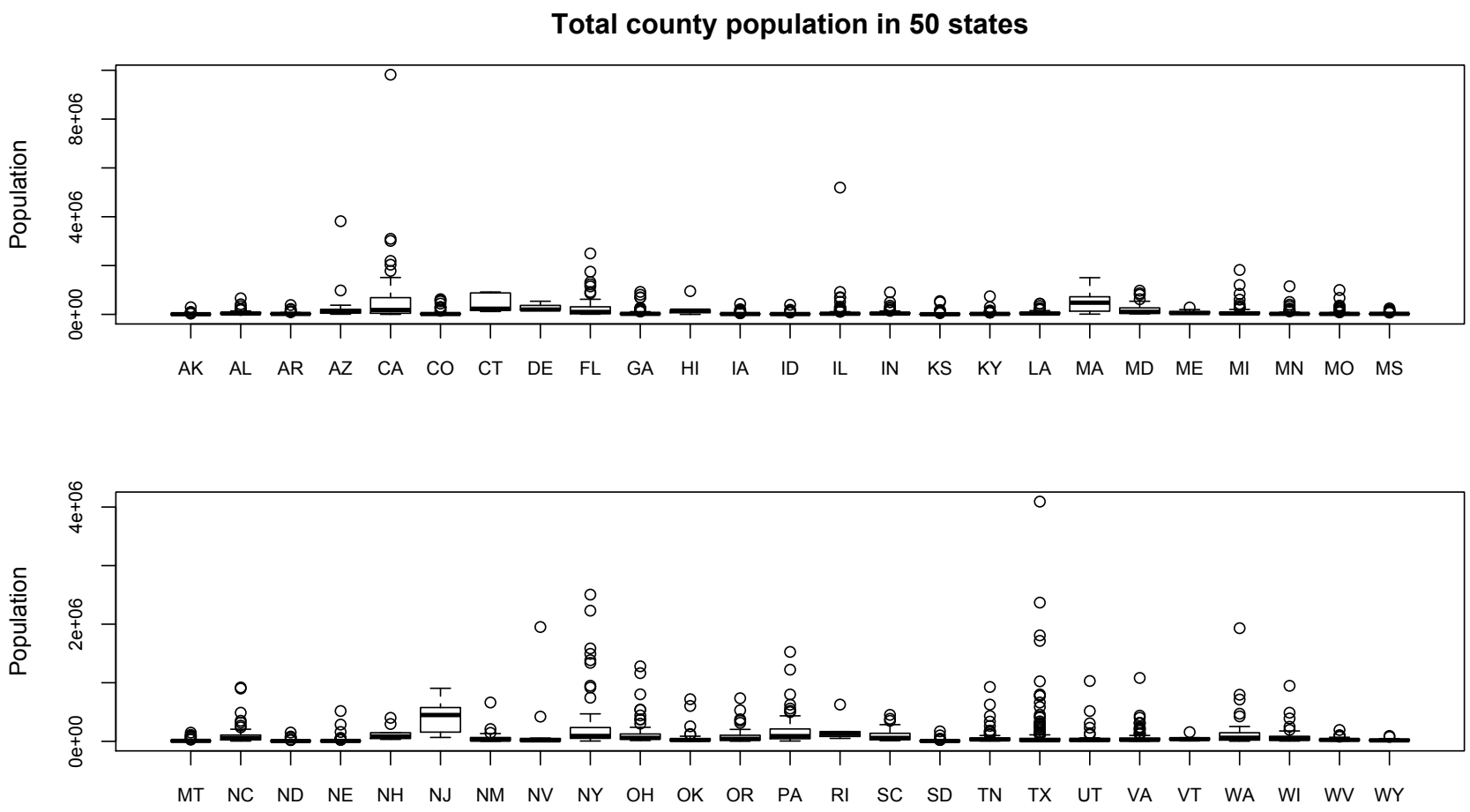
Ch4. Transforming data

1. Motivation: Examples of re-expressed batches
2. What, how, why?
3. Family of re-expressions
4. Why did re-expressions work?
5. Diagnostic plots

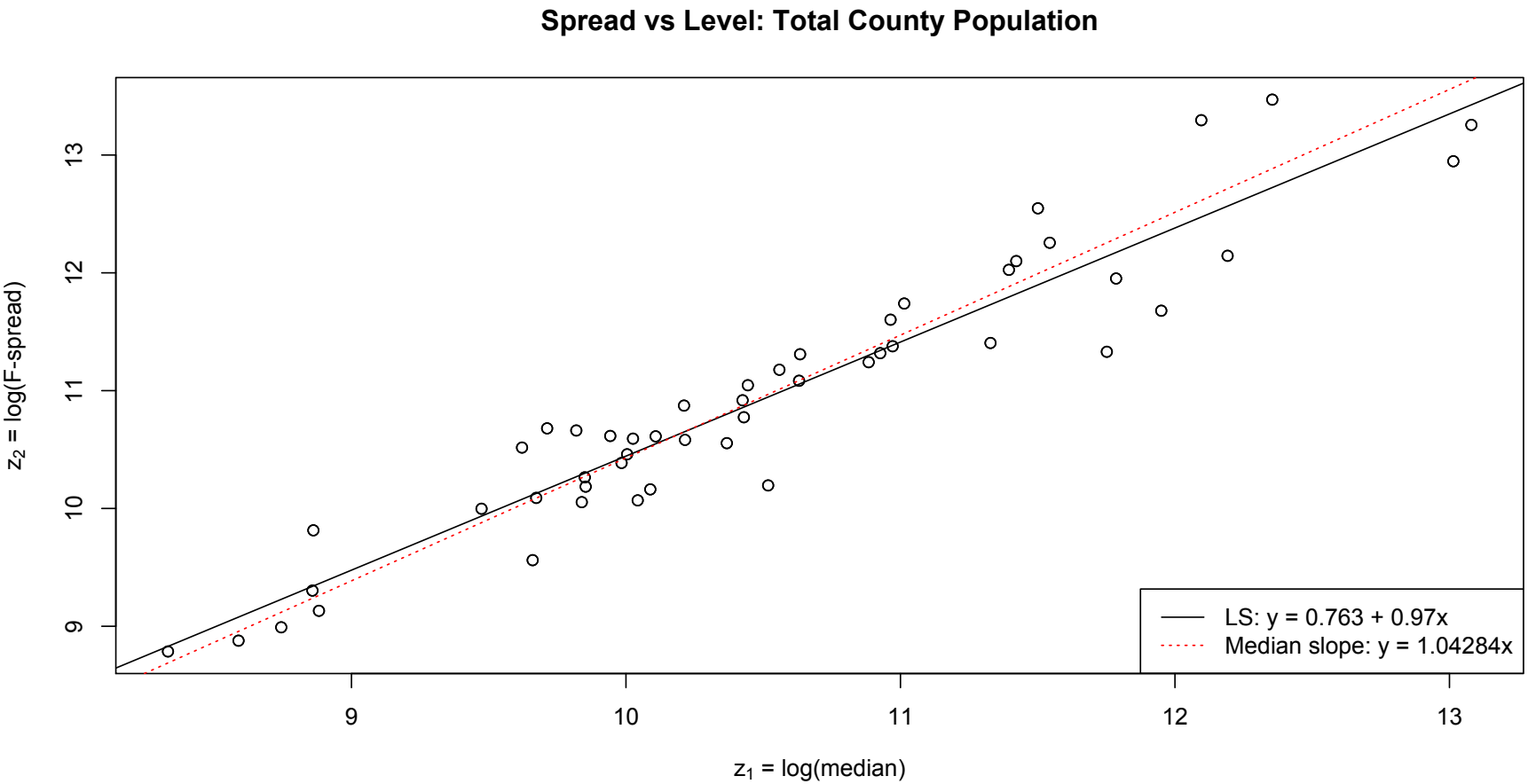
Transforming data



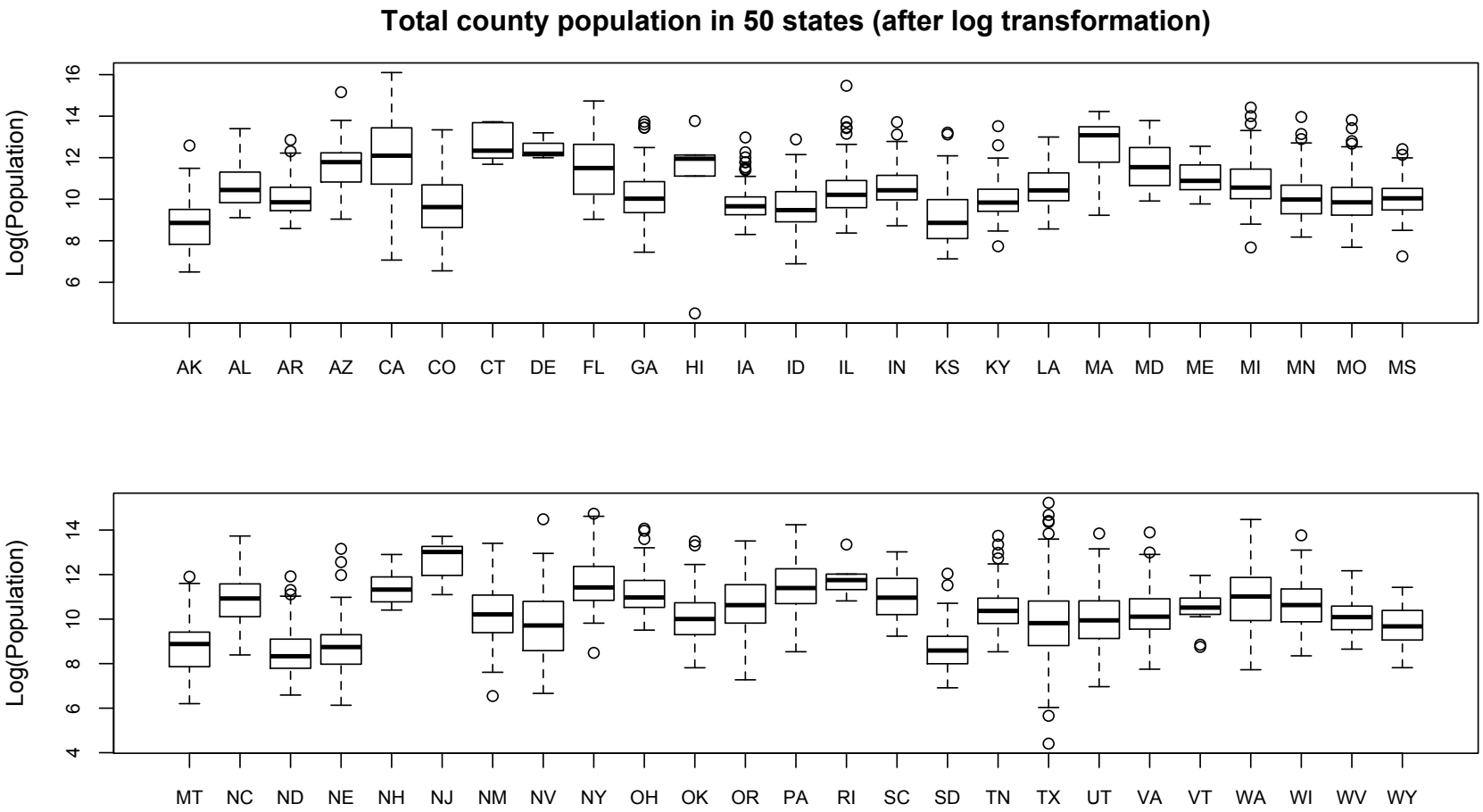
Transforming data



Transforming data



Transforming data



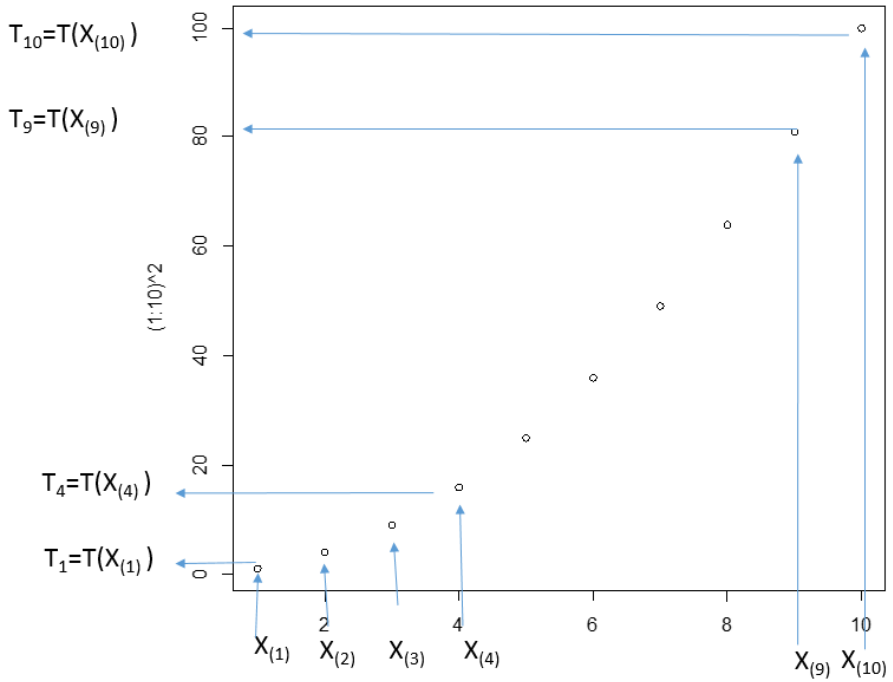
What is transformation/re-expression?

Apply a single mathematical function T to all raw data values:

Replace $x_1, \ldots, x_n$ by $T(x_1), \ldots, T(x_n)$

Monotonic Transforms

Want Monotonic Increasing Transforms



If T(x) is monotonic and smooth then order is preserved under the transform.

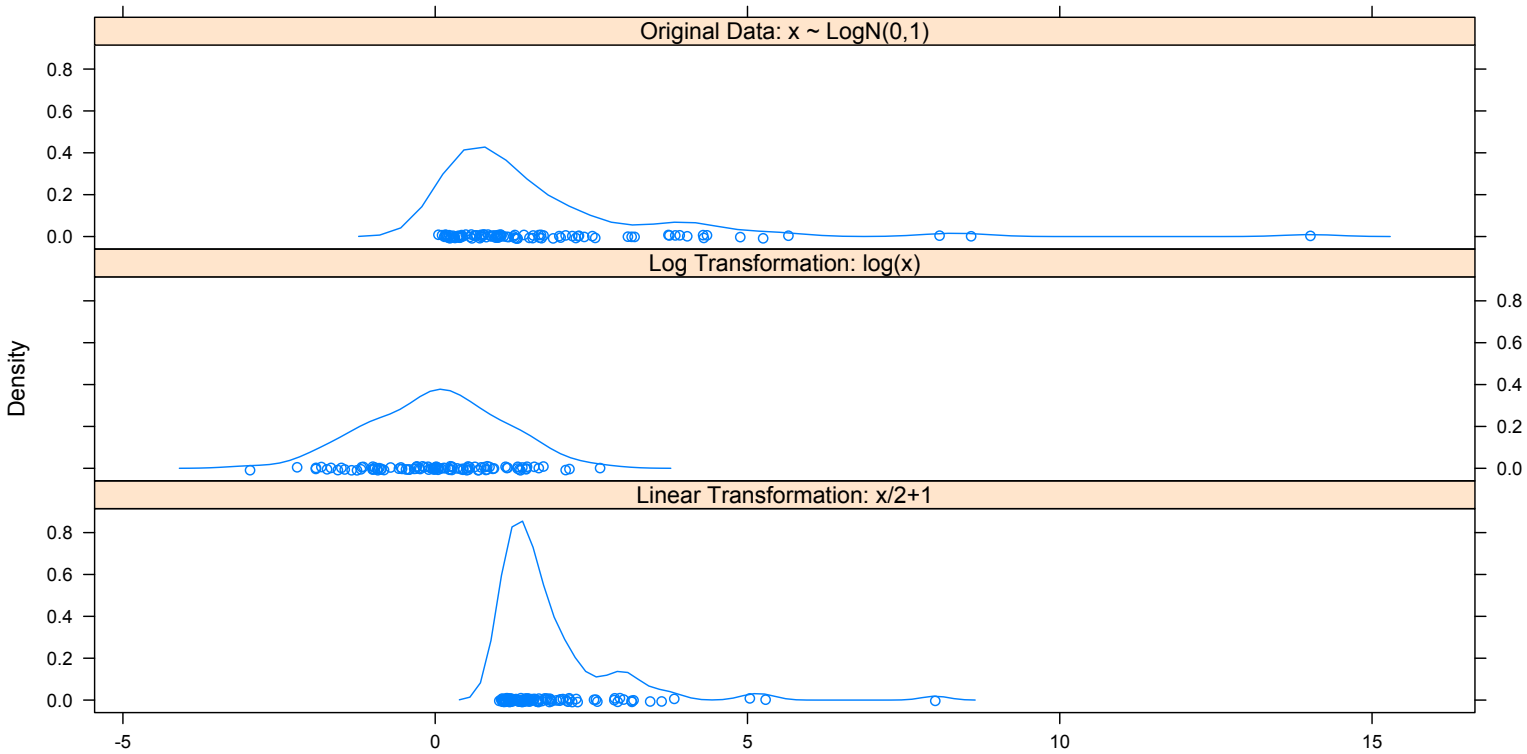
Median(T(Data)) = T(Median(Data))

Upper Fourth(T(Data)) = T (Upper Fourth(Data))

Lower Fourth(T(Data)) = T (Lower Fourth(Data))

How does transformation change the shape of data?

1. No change in shape (linear transformation): change origin and scale
2. Change shape: stronger transformations such as logarithm and square root



Reasons to transform data:

- Facilitate interpretation (temperature ($^\circ$F, $^\circ$C); population: Exp function$\rightarrow$log)
- Achieve symmetry in a batch (why: estimate of location)
- Promote more stable spread across groups
- Achieve more linear relationship or straightness between two variables (y-versus-x)
- Simplify structure for two-way table /additive structure
- Many outliers in one tail

Improvement after transformation

- Informative display
- Effective summary
- Uncomplicated analysis
- Enhance interpretation

Objectives

1. What set (family) of re-expression?
   – Define a convenient family of transformations
   – convenient $\equiv$ monotonic; e.g. $T(x_M) = T(x)_M$
2. Which re-expression (member of family) to choose for single batch?
3. Which re-expression for multiple batches?

Power transformations

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

where a, b, c, d, and p are real numbers. Properties:

- Require $a > 0$ for $p > 0$, $a < 0$ for $p < 0$, so $T_p(x)$ preserves the order of data (thus, preserve letter values)
- Continuous, smooth functions
- Concavity. Compress the scale for larger data value (or smaller data values) more than it does for smaller ones (or larger ones)
- Flexibility. Restrict to ladder transformations
- Geometric unity

Ladder of Transformations: $y = x^p$

$p = 2$	x^2	square
$p = 1$	x	(None)
$p = \frac{1}{2}$	$\sqrt{x}$	square root
$p = 0$	$\log(x)$	logarithm
$p = -\frac{1}{2}$	$1/\sqrt{x}$	reciprocal square root
$p = -1$	$1/x$	reciprocal
$p = -2$	$1/x^2$	reciprocal square

Choice of a, b, c, d: Compare transformations

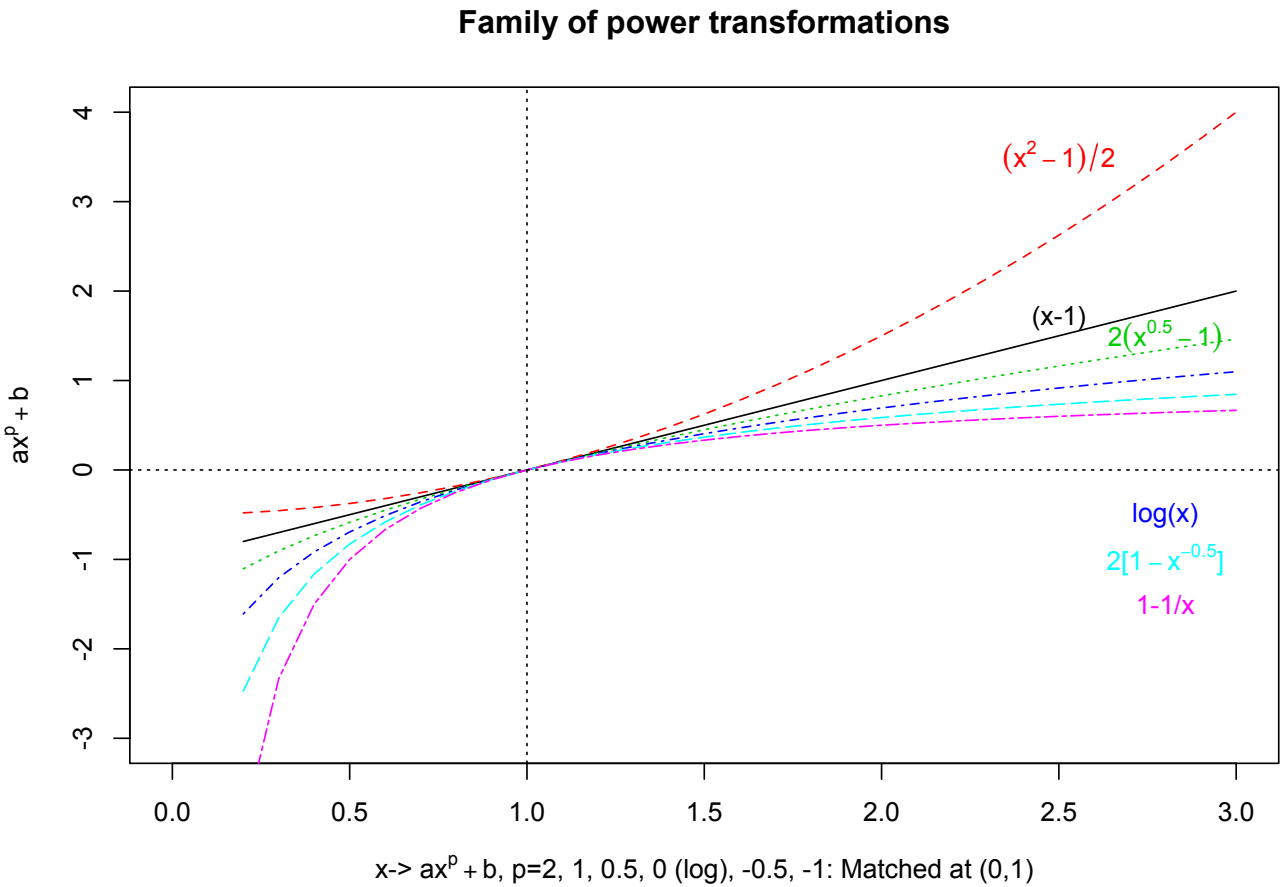
$$T_p(x) = \begin{cases} \frac{x^p - 1}{p} & p \neq 0 \\ \ln x & p = 0 \end{cases} \qquad (\lim_{p \to 0} \frac{x^p - 1}{p} = \ln x^1)$$

Properties:

- Curves share common point, common slope at that point $T_p(x = 1) = 0,\ T'_p(x = 1) = 1$.
- Curves pass through the common point in the same direction
- Curves are ordered by p
- Closeness of functions
- log transformation fits between $T^*_{1/2}(x)$ and $T^*_{-1/2}(x)$ or $T^*_{\lambda}(x)$ and $T^*_{-\lambda}(x)$, $\forall \lambda > 0$.

[1] $\ln x \approx 2.303 \log_{10} x$

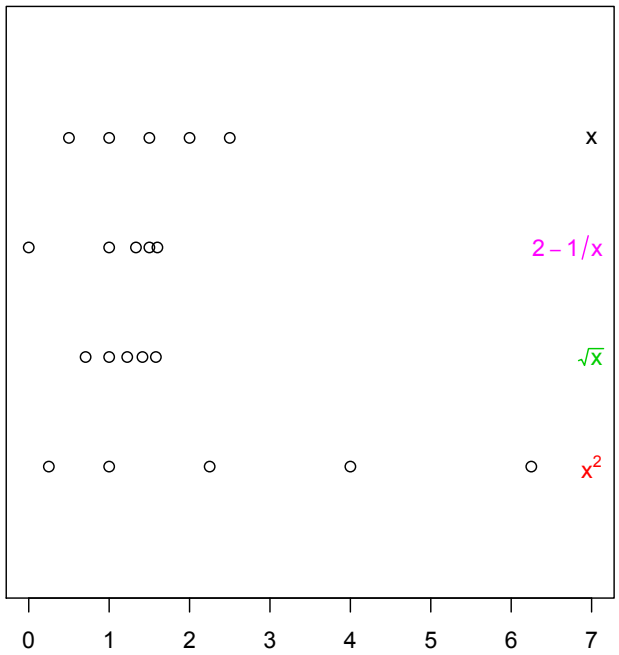
Power transformations

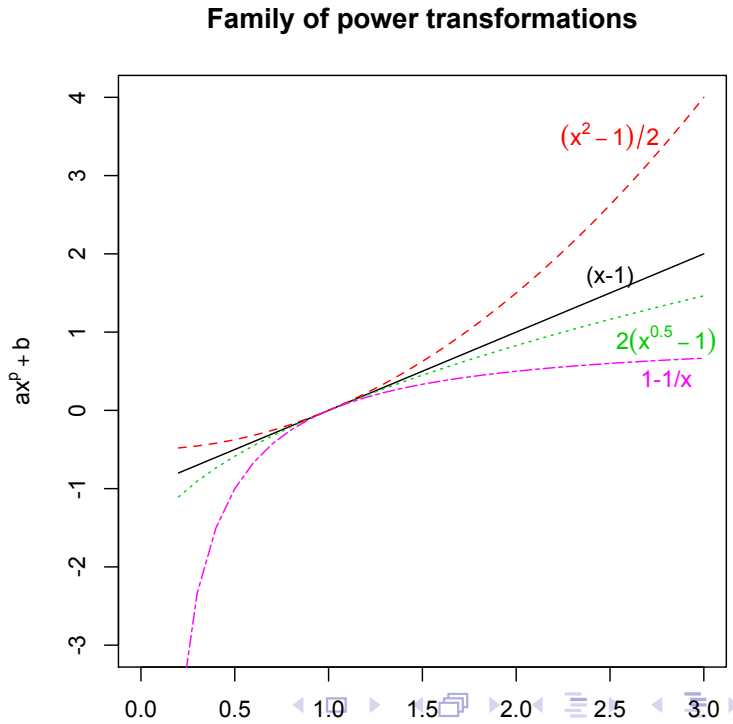


$T_p(x)$ all look about the same near middle $(1,0)$.

Effect of power transformations

- Transform x to $T_p(x) \equiv x^p$: Changes shape of x
- Larger $p \Rightarrow$ Greater effect
- $T_p(x) = x^p \Rightarrow T_p^*(x) \equiv a + bx^p$ does not change shape!
- Effect of p in $T_p(x)$ or $T_p^*(x)$ more noticeable as x moves away from center





How do we choose p? – Re-expression for symmetry

- Example: Infant mortality data, 1997, 100 counties in North Carolina

How do we choose p? – Re-expression for symmetry

- Example: Infant mortality data, 1997, 100 counties in North Carolina

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	50.5	5	5.0	5.00	0.0	0.0000
F	25.5	2	11.0	6.50	9.0	6.6717
E	13.0	1	18.0	9.50	17.0	7.3891
D	7.0	0	30.0	15.00	30.0	9.7776
C	4.0	0	58.0	29.00	58.0	15.5685
B	2.5	0	67.5	33.75	67.5	15.6694
A	1.5	0	74.0	37.00	74.0	15.3047
Z	1.0	0	76.0	38.00	76.0	14.2854

- Check symmetry: a set of midsummaries, one for each pair of letter values

How do we choose p? – Re-expression for symmetry

- Check symmetry: a set of midsummaries, one for each pair of letter values
- If no transformation is needed, both "mids" and "p-Sigma" should be (??)

How do we choose p? – Re-expression for symmetry

- Check symmetry: a set of midsummaries, one for each pair of letter values
- If no transformation is needed, both "mids" and "p-Sigma" should be (??)
  - Perfectly symmetric batch: all midsummaries equal to the median
  - Skew to the right: midsummaries increase
  - Skew to the left: midsummaries decrease

How do we choose p? – Re-expression for symmetry

- Transformation plot for symmetry:

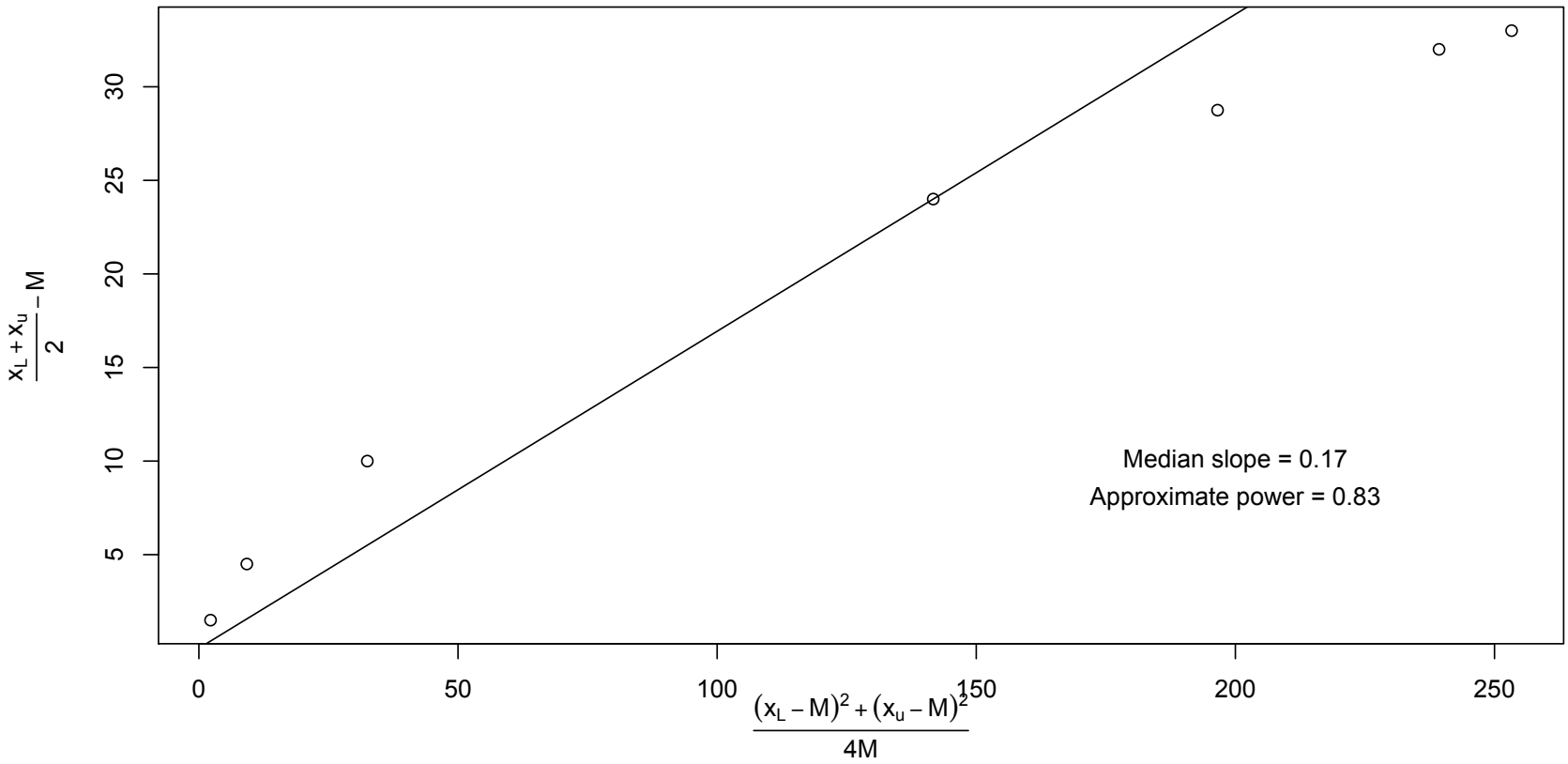
M=median, x_L, x_U = lower, upper letter values,
d_L, d_U=distance from x_L and x_U to M.

1. $$X = [(x_U - M)^2 + (x_L - M)^2]/(4M) = (d_U^2 + d_L^2)/(4M)$$

$$Y = (x_L + x_U)/2 - M = (d_U - d_L)/2$$

2. Plot Y versus X: slope $= 1 - p = 1-$ transformation power

Example


$\frac{x_L + x_u}{2} - M$
5
10
15
20
25
30
0
50
100
150
200
250
$\frac{(x_L - M)^2 + (x_u - M)^2}{4M}$
Median slope = 0.17
Approximate power = 0.83

How do we choose p? – Re-expression for symmetry

After transformation (Infant Mortality)$^{0.83}$

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	50.5	3.8032	3.8032	3.8032	0.0000	0.0000
F	25.5	1.7777	7.3174	4.5475	5.5397	4.1066
E	13.0	1.0000	11.0123	6.0061	10.0123	4.3518
D	7.0	0.0000	16.8272	8.4136	16.8272	5.4843
C	4.0	0.0000	29.0834	14.5417	29.0834	7.8067
B	2.5	0.0000	32.9751	16.4875	32.9751	7.6548
A	1.5	0.0000	35.5992	17.7996	35.5992	7.3626
Z	1.0	0.0000	36.3978	18.1989	36.3978	6.8415

Why it works? (Ch8)

Suppose data x^p are symmetric, then

$$x_U^p - M^p = M^p - x_L^p \Rightarrow (x_U^p + x_L^p)/2 = M^p$$

If $p \neq 0$, Taylor series expansion of x^p about M:

$$x_U^p \approx M^p + pM^{p-1}(x_U - M) + \frac{p(p-1)}{2}M^{p-2}(x_U - M)^2$$

$$x_L^p \approx M^p + pM^{p-1}(x_L - M) + \frac{p(p-1)}{2}M^{p-2}(x_L - M)^2$$

$$\frac{x_U^p + x_L^p}{2} \approx \frac{1}{2}\left\{2M^p + pM^{p-1}(x_U + x_L - 2M) + \frac{p(p-1)}{2}M^{p-2}[(x_U - M)^2 + (x_L - M)\right.$$

$$\Rightarrow M(x_U + x_L - 2M) + \frac{p-1}{2}[(x_U - M)^2 + (x_L - M)^2] \approx 0$$

$$\Rightarrow \frac{x_U + x_L}{2} - M \approx (1-p)\frac{[(x_U - M)^2 + (M - x_L)^2]}{4M}$$

Why it works? (Ch8)

- If slope $= 0$, $p = 1$, no need transformation
- Skew to the right, mid increases, $1 - p > 0$, $p < 1$ (pull in the right tail);
- midsummaries increase, transform to lower powers

Example

County population in 2010:

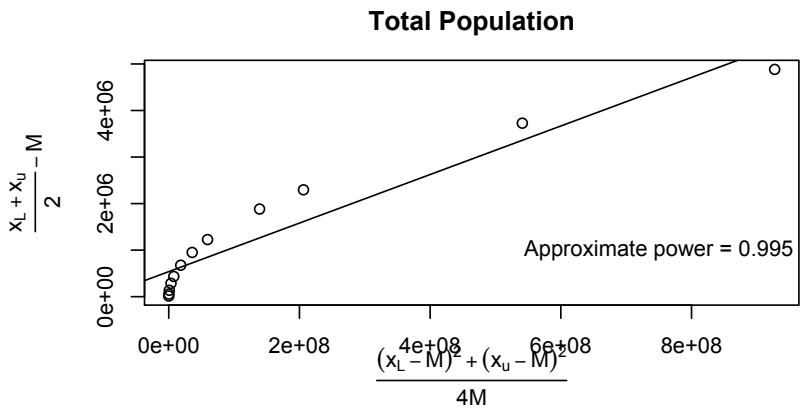
	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	1572.0	25857.0	25857.0	25857.00	0.0	0.00
F	786.5	11104.5	66699.0	38901.75	55594.5	41212.26
E	393.5	6157.0	157906.5	82031.75	151749.5	65958.00
D	197.0	3423.0	321520.0	162471.50	318097.0	103674.06
C	99.0	2071.0	622263.0	312167.00	620192.0	166473.77
B	50.0	1321.0	919040.0	460180.50	917719.0	213039.09
A	25.5	813.0	1401948.0	701380.50	1401135.0	289783.00
Z	13.0	662.0	1951269.0	975965.50	1950607.0	366646.15
Y	7.0	494.0	2504700.0	1252597.00	2504206.0	433909.01
X	4.0	416.0	3817117.0	1908766.50	3816701.0	616139.72
W	2.5	188.0	4643567.0	2321877.50	4643379.0	704141.15
V	1.5	86.0	7506640.0	3753363.00	7506554.0	1076330.64
U	1.0	82.0	9818605.0	4909343.50	9818523.0	1338282.67

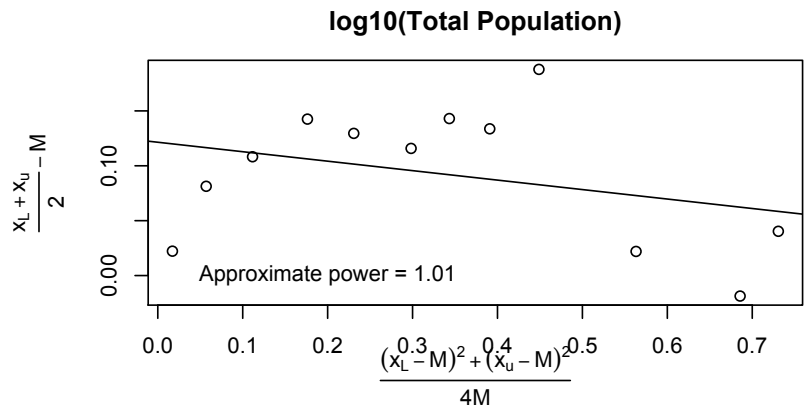
Example

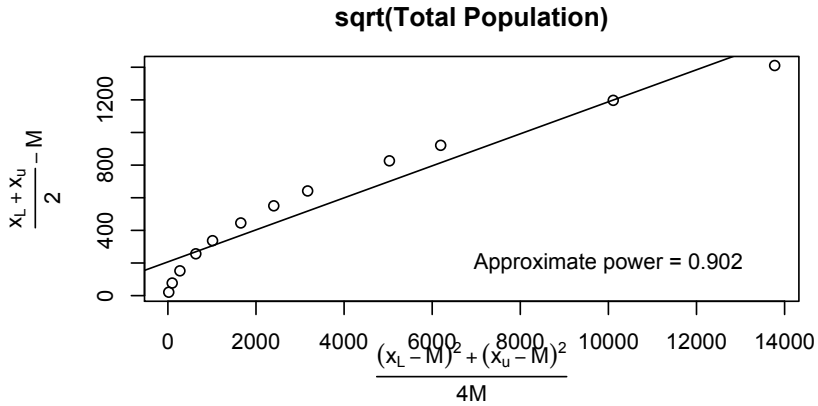
County population in 2010:

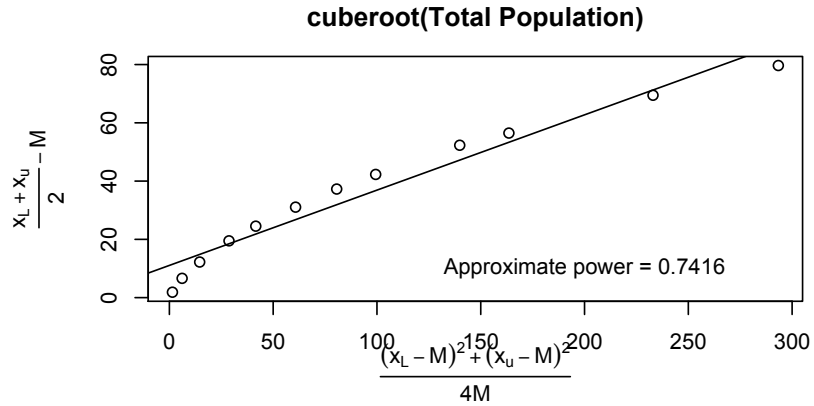
	Depth	Mid	Spread	pseudo-s
M	1572.0	25857.00	0.0	0.00
F	786.5	38901.75	55594.5	41212.26
E	393.5	82031.75	151749.5	65958.00
D	197.0	162471.50	318097.0	103674.06
C	99.0	312167.00	620192.0	166473.77
B	50.0	460180.50	917719.0	213039.09
A	25.5	701380.50	1401135.0	289783.00
Z	13.0	975965.50	1950607.0	366646.15
Y	7.0	1252597.00	2504206.0	433909.01
X	4.0	1908766.50	3816701.0	616139.72
W	2.5	2321877.50	4643379.0	704141.15
V	1.5	3753363.00	7506554.0	1076330.64
U	1.0	4909343.50	9818523.0	1338282.67

Example: County populations







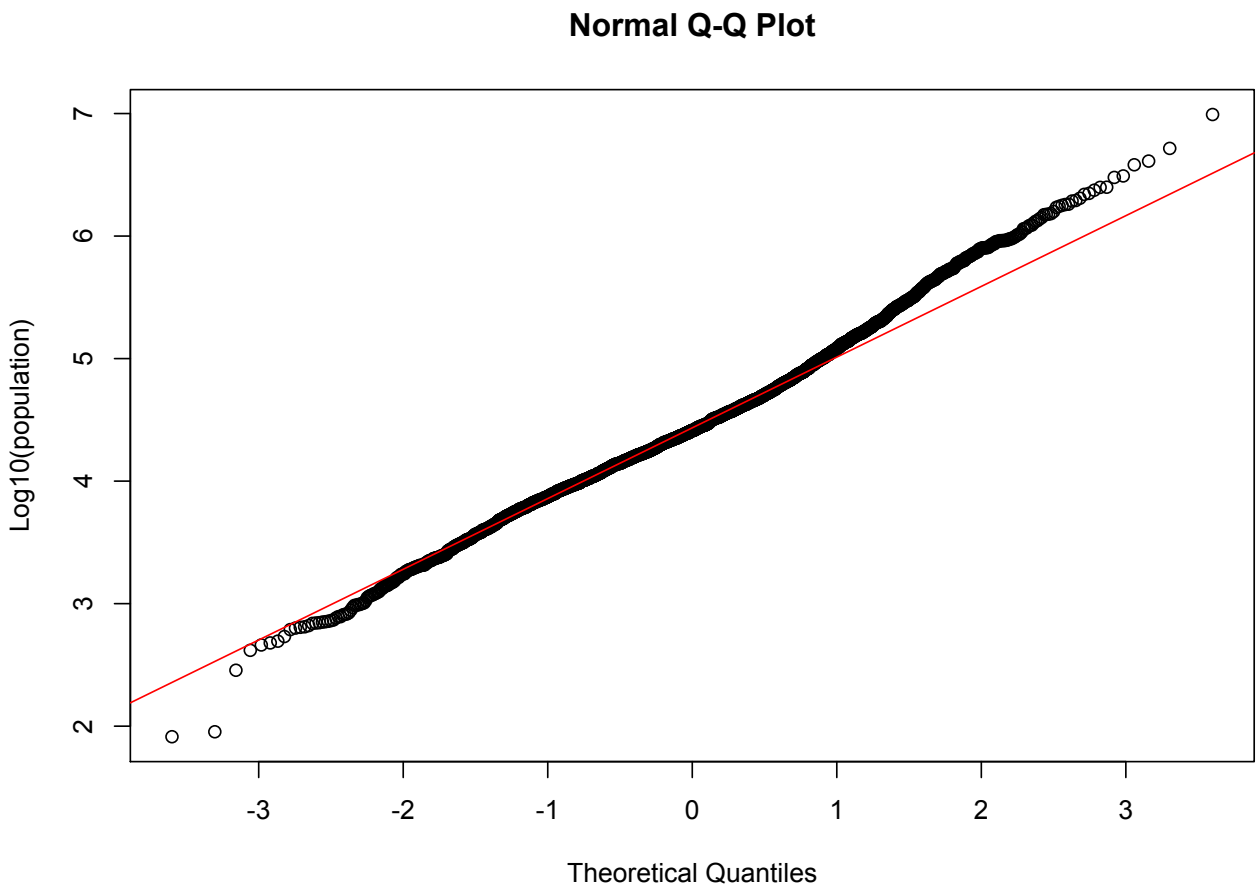


Example

Log10 County population:

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	1572.0	4.4126	4.4126	4.4126	0.0000	0.0000
F	786.5	4.0455	4.8241	4.4348	0.7786	0.5772
E	393.5	3.7894	5.1984	4.4939	1.4090	0.6124
D	197.0	3.5344	5.5072	4.5208	1.9728	0.6430
C	99.0	3.3162	5.7940	4.5551	2.4778	0.6651
B	50.0	3.1209	5.9633	4.5421	2.8424	0.6598
A	25.5	2.9101	6.1467	4.5284	3.2366	0.6694
Z	13.0	2.8209	6.2903	4.5556	3.4695	0.6521
Y	7.0	2.6937	6.3988	4.5462	3.7050	0.6420
X	4.0	2.6191	6.5817	4.6004	3.9626	0.6397
W	2.5	2.2053	6.6638	4.4345	4.4585	0.6761
V	1.5	1.9340	6.8538	4.3939	4.9198	0.7054
U	1.0	1.9138	6.9920	4.4529	5.0782	0.6922

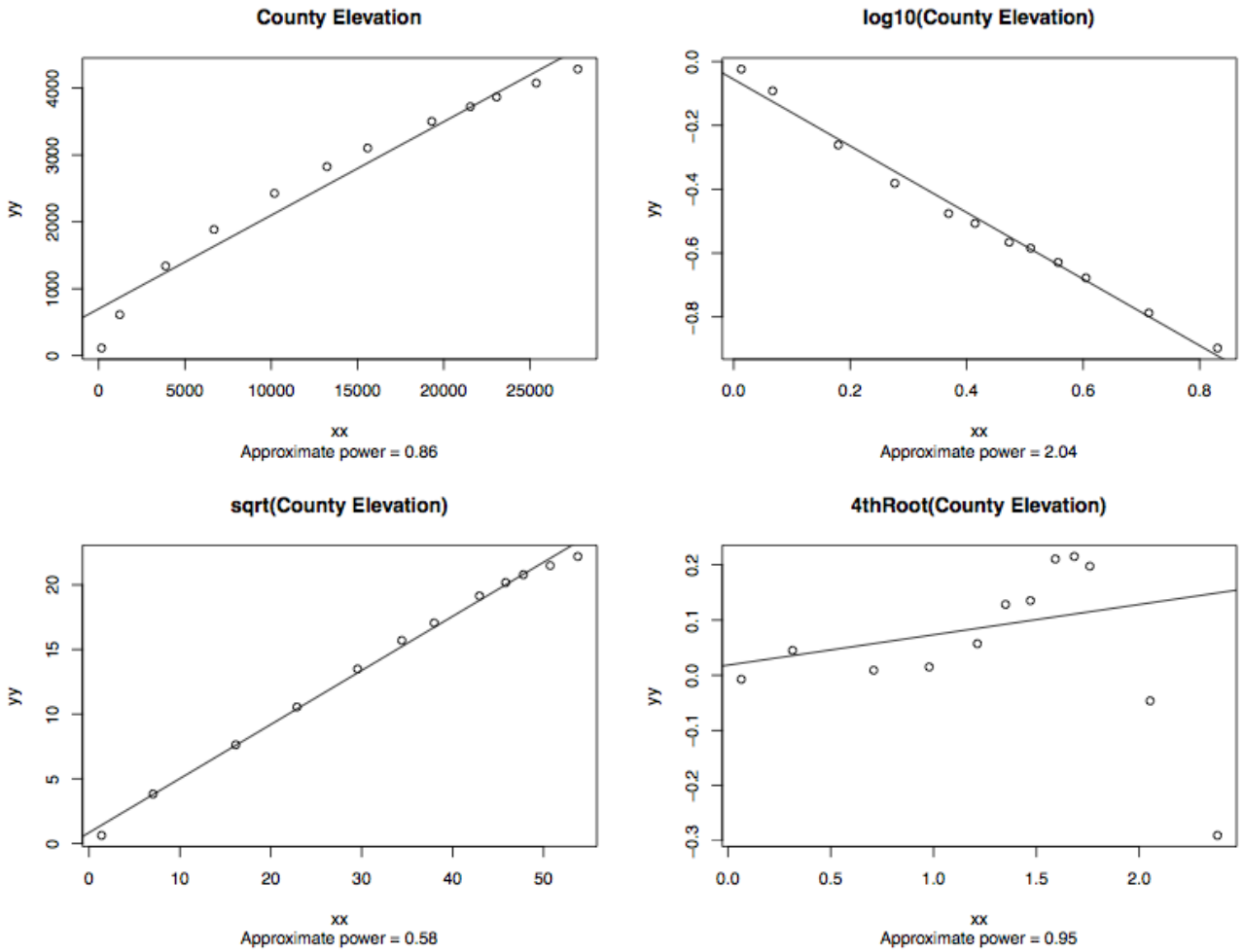
Example: County populations



Example: County elevations

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	1534.5	795.0	795.0	795.00	0.0	0.00
F	767.5	400.0	1415.0	907.50	1015.0	752.42
E	384.0	155.0	2660.0	1407.50	2505.0	1088.80
D	192.5	44.0	4226.0	2135.00	4182.0	1363.00
C	96.5	19.5	5340.5	2680.00	5321.0	1428.28
B	48.5	10.0	6433.5	3221.75	6423.5	1491.15
A	24.5	7.5	7234.0	3620.75	7226.5	1494.59
Z	12.5	5.0	7791.5	3898.25	7786.5	1463.59
Y	6.5	4.0	8590.0	4297.00	8586.0	1487.71
X	3.5	3.0	9033.0	4518.00	9030.0	1457.74
W	2.0	2.0	9322.0	4662.00	9320.0	1413.32
V	1.5	1.0	9740.0	4870.50	9739.0	1396.43
U	1.0	0.0	10158.0	5079.00	10158.0	1384.55

Example: County elevations



Example: County elevations

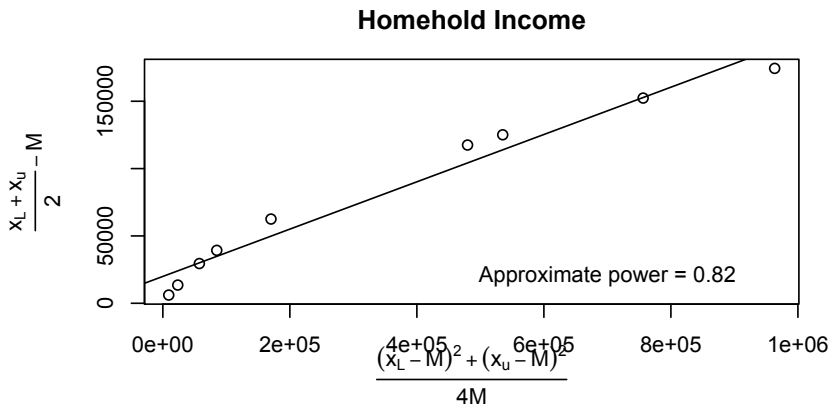
Fourth roots:

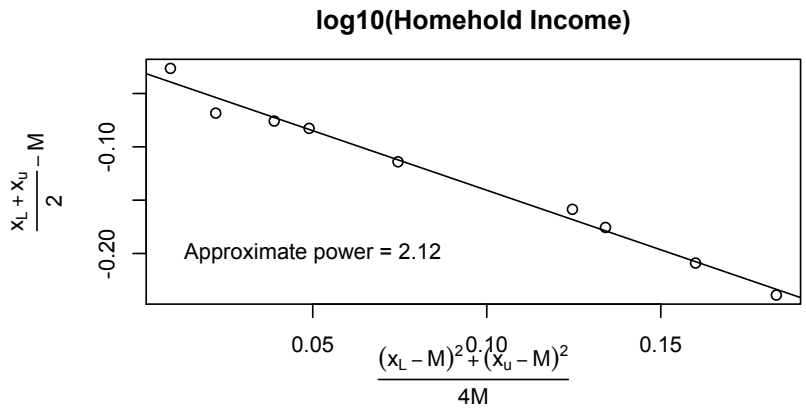
	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	1534.5	5.3100	5.3100	5.3100	0.0000	0.0000
F	767.5	4.4721	6.1332	5.3027	1.6611	1.2314
E	384.0	3.5284	7.1816	5.3550	3.6531	1.5878
D	192.5	2.5754	8.0627	5.3191	5.4873	1.7884
C	96.5	2.1013	8.5486	5.3249	6.4473	1.7306
B	48.5	1.7783	8.9560	5.3671	7.1777	1.6662
A	24.5	1.6542	9.2224	5.4383	7.5682	1.5653
Z	12.5	1.4953	9.3951	5.4452	7.8998	1.4849
Y	6.5	1.4142	9.6271	5.5206	8.2128	1.4231
X	3.5	1.3017	9.7489	5.5253	8.4472	1.3637
W	2.0	1.1892	9.8260	5.5076	8.6368	1.3097
V	1.5	0.5946	9.9326	5.2636	9.3380	1.3389
U	1.0	0.0000	10.0393	5.0196	10.0393	1.3684

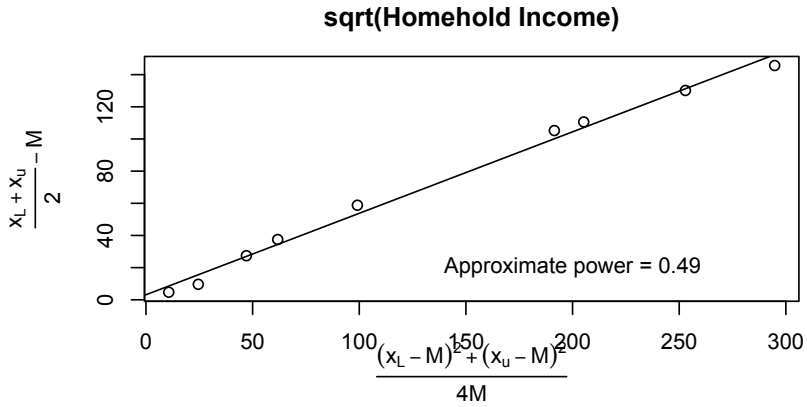
Example: Household Income

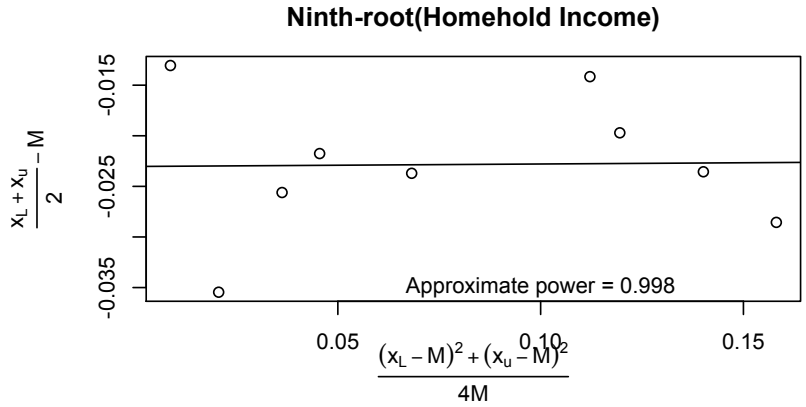
	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	39200	39200	39200	0	0.00
F	98.0	19000	71600	45300	52600	38992.44
E	49.5	12000	93450	52725	81450	35402.29
D	25.0	8400	129000	68700	120600	39305.91
C	13.0	7000	150000	78500	143000	38384.48
B	7.0	4570	198900	101735	194330	45111.72
A	4.0	2380	311000	156690	308620	63828.85
Z	2.5	2100	326500	164300	324400	60975.90
Y	1.5	1600	381500	191550	379900	65826.07
X	1.0	1200	426000	213600	424800	68576.54

Example: Household Income









Example: Household Income

log10(Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	4.5933	4.5933	4.5933	0.0000	0.0000
F	98.0	4.2788	4.8549	4.5668	0.5762	0.4271
E	49.5	4.0792	4.9706	4.5249	0.8914	0.3874
D	25.0	3.9243	5.1106	4.5174	1.1863	0.3866
C	13.0	3.8451	5.1761	4.5106	1.3310	0.3573
B	7.0	3.6599	5.2986	4.4793	1.6387	0.3804
A	4.0	3.3766	5.4928	4.4347	2.1162	0.4377
Z	2.5	3.3217	5.5137	4.4177	2.1919	0.4120
Y	1.5	3.1901	5.5785	4.3843	2.3884	0.4138
X	1.0	3.0792	5.6294	4.3543	2.5502	0.4117

Example: Household Income

Square-root (Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	197.9899	197.9899	197.9899	0.0000	0.0000
F	98.0	137.8405	267.5818	202.7111	129.7413	96.1774
E	49.5	109.5445	305.6957	207.6201	196.1511	85.2572
D	25.0	91.6515	359.1657	225.4086	267.5142	87.1881
C	13.0	83.6660	387.2983	235.4822	303.6323	81.5019
B	7.0	67.6018	445.9821	256.7919	378.3803	87.8371
A	4.0	48.7852	557.6737	303.2295	508.8885	105.2484
Z	2.5	45.8128	571.3279	308.5703	525.5151	98.7785
Y	1.5	39.6812	616.6019	328.1415	576.9207	99.9643
X	1.0	34.6410	652.6868	343.6639	618.0457	99.7727

Example: Household Income

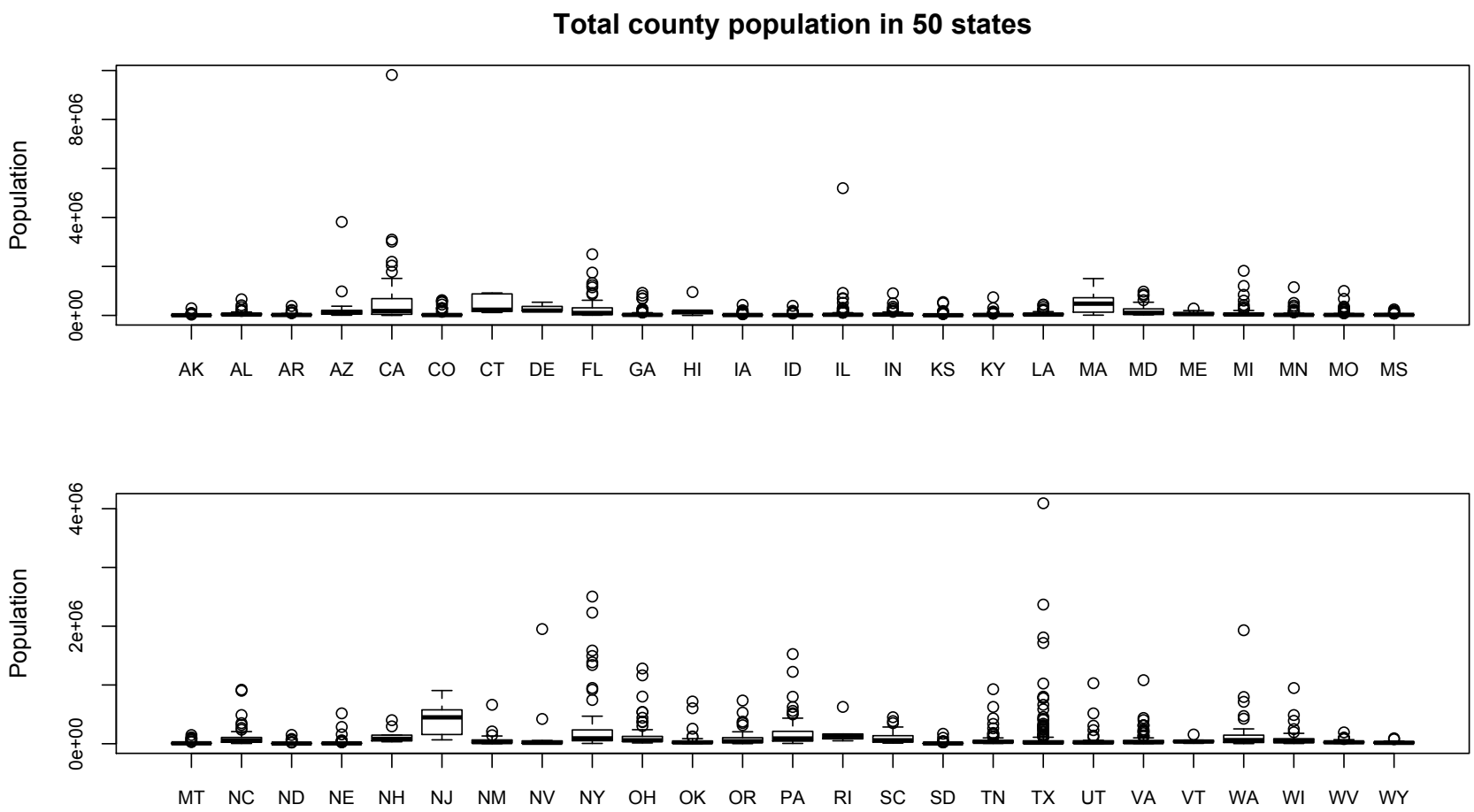
Ninth-root (Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	3.2387	3.2387	3.2387	0.0000	0.0000
F	98.0	2.9883	3.4629	3.2256	0.4746	0.3518
E	49.5	2.8395	3.5669	3.2032	0.7274	0.3161
D	25.0	2.7292	3.6969	3.2131	0.9678	0.3154
C	13.0	2.6744	3.7594	3.2169	1.0850	0.2912
B	7.0	2.5507	3.8792	3.2149	1.3285	0.3084
A	4.0	2.3723	4.0767	3.2245	1.7044	0.3525
Z	2.5	2.3393	4.0986	3.2189	1.7593	0.3307
Y	1.5	2.2627	4.1675	3.2151	1.9047	0.3300
X	1.0	2.1985	4.2217	3.2101	2.0232	0.3266

Multiple batches: How to stabilize spread?

- When data are amounts or counts, often find a systematic relationship between spread and level: increasing level usually brings increasing spread.
- Advantage of transformed data (no dependence of spread on level)
  - Better suited for comparison and visual exploration
  - Better suited for common confirmatory techniques (ANOVA)
  - Individual batches may become more nearly symmetric, fewer outliers.

Stabalize Spread



Multiple batches: How to stabilize spread?

- When to transform data in multiple batches?
  – Batches comparison show a systematic relationship between spread and level
  – Want to eliminate dependency for better visual exploration or future analysis, e.g., ANOVA
- What is the transformation for multiple batches?
  – Replace x by x^p
  – $p = 0$, use $\log x$
- How to find an appropriate power transformation: Spread-vs-level plot
  1. For each batch, calculate median & F-spread: Med_j, FS_j
  2. Plot $\log(FS_j)$ (y-axis), $\log(Med_j)$ (x-axis)
  3. If plot has zero slope, no transform needed
  4. If plot has slope s, use $p = 1 - s$.

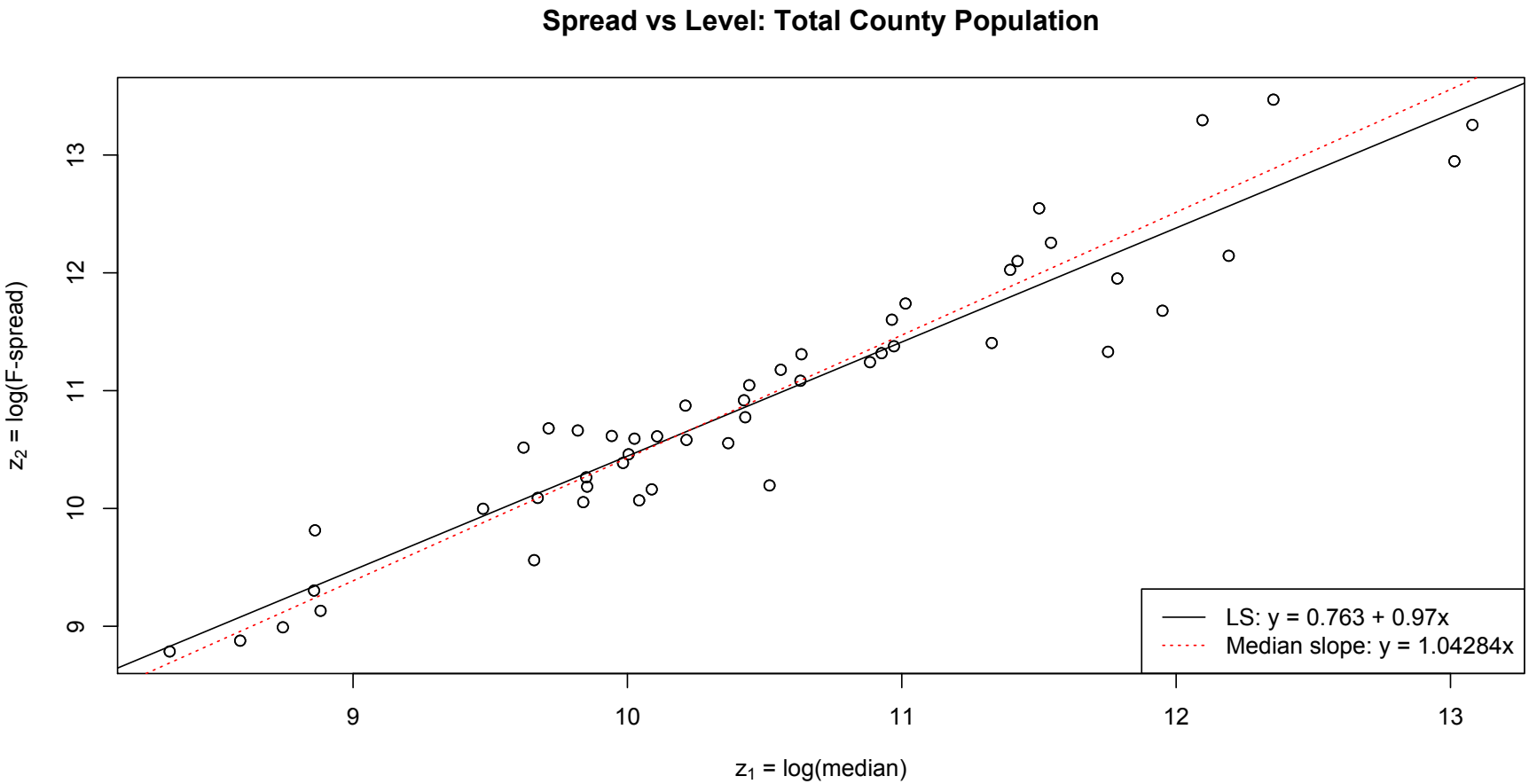
Multiple batches: How to stabilize spread?

Construction of spread-versus-level plots

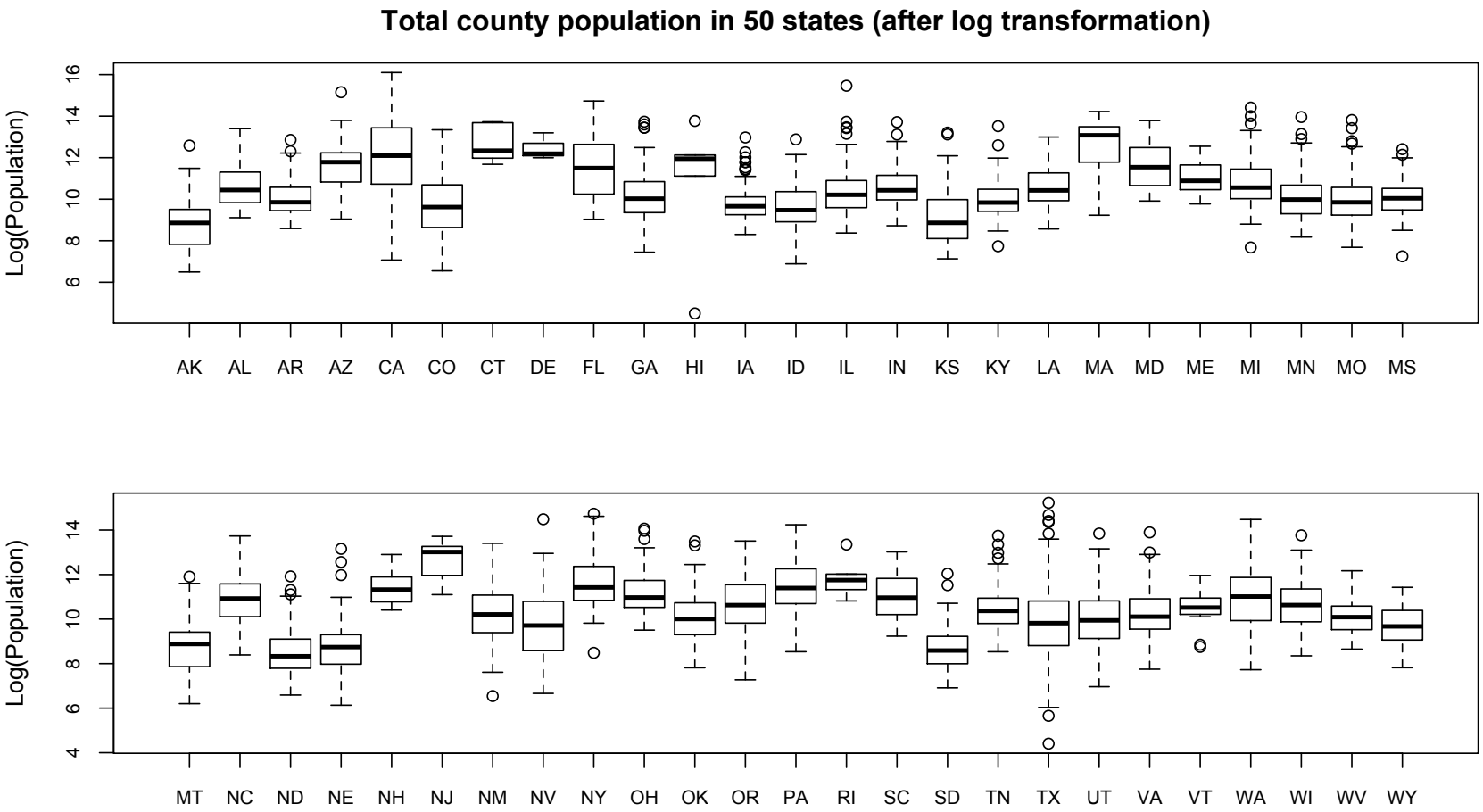
- Understand relationship between spread and level
  Suppose $d_F = CM^b$, $\log d_F = \log C + b \log M = k + b \log M$
- Spread-versus-level plot: a plot of $\log d_F$ against $\log M$ for all batches
- b is slope, $p = 1 - b$ is the approximate value of exponent for a power transformation of x to stabilize spread. $b = 1$, $p = 0$, use logarithms. $b = 1/2$, $p = 1/2$, square root transformations. (simplicity, interpretability)
- Recheck with (d_F vs M) plot
- Make choices:

  (1) equalize spread;
  (2) subject-matter explanation.
  demography (log $\rightarrow$ linear): (advantages of linear growth: interpretability, ease of detecting departure from fit, convenience in interpolation);
  (3) choose a power that's an integer multiple of $1/2$

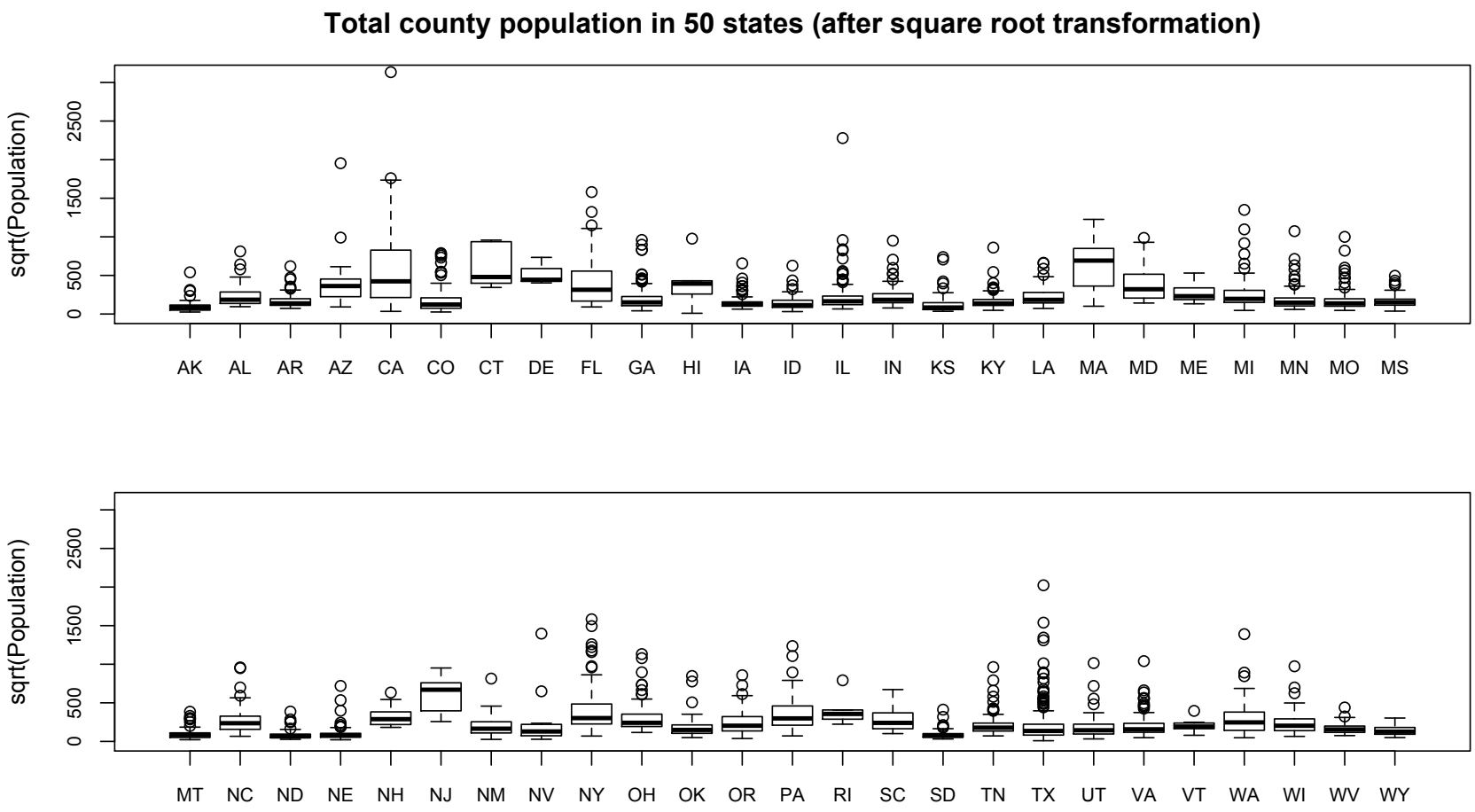
How to stabilize spread?



How to stabilize spread?



How to stabilize spread?



Spread-vs-Level Plot: Why it works? (Ch3, Ch8)

- Raw data: X, transformed data: x^p.
- Suppose for each batch, x^p is the "right" transformation, then fourth-spread of x^p is roughly constant.

	x^p	x
Median	m	$m^{1/p} \equiv m^q$
UF	$m + d$	$(m + d)^{1/p}$
LF	$m - c$	$(m - c)^{1/p}$
F-spread	$d + c$	$(m + d)^{1/p} - (m - c)^{1/p}$

- Taylor series:
  $f(z) \approx f(z_0) + f'(z_0)(z - z_0) + f''(z_0)(z - z_0)^2/2$. When $f(z) = (1 + z)^q$, and $z_0 = 0$,

$$(1 + z)^q \approx 1 + qz + [q(q - 1)/2]z^2$$

Spread-vs-Level Plot

Denote $q \equiv 1/p$, $y \equiv d/m$, $z \equiv c/m$. Then F-spread for "wrong" scale (x) is:

$$
\begin{aligned}
& (m+d)^q - (m-c)^q \\
=& m^q[(1+d/m)^q - (1-c/m)^q] \\
=& m^q[(1+y)^q - (1-z)^q] \\
\approx& m^q[1+yq+y^2q(q-1)/2 - 1 + zq - z^2q(q-1)/2] \\
=& m^q(y+z)q[1+(y-z)(q-1)/2+\cdots] \\
=& m^q[(d+c)q/m][1+(d-c)(q-1)/(2m)+\cdots] \\
=& m^{q-1}[(d+c)q][1+(d-c)(q-1)/(2m)+\cdots]
\end{aligned}
$$

Spread-vs-Level Plot

Take log of both sides:

$$\log(F - \mathit{spread}) \approx \log(q) + \log(d + c) + (q - 1)\log(m) + \log(1 + \mathit{tiny}) \quad (1)$$

$$\approx \log(q) + \log(d + c) + (q - 1)\log(m^{1/p})p \quad (2)$$

$$= \log(q) + \log(d + c) + [(q - 1)/(1/p)]\log(\text{median of x} \quad (3)$$

$$= \log(q) + \log(d + c) + (1 - p)\log(\text{median of x}) \quad (4)$$

Plot $\log(F - s)$ **vs.** $\log(med)$**: slope =** $1 - p$**.**

Note: assume c, d small, $p \neq 0$.

Spread-vs-Level Plot

For p = 0 (separate algebra):

	$\log 10(x)$	x
median	m	10^m
UF	$m + d$	10^{m+d}
LF	$m - c$	10^{m-c}
FS	$d + c$	$10^m(10^d - 10^c)$

$$\log 10(\text{FS of x}) = m + \log 10(10^d - 10^c)$$

$$= \log 10(\text{median of x}) + \log 10(10^d - 10^c)$$

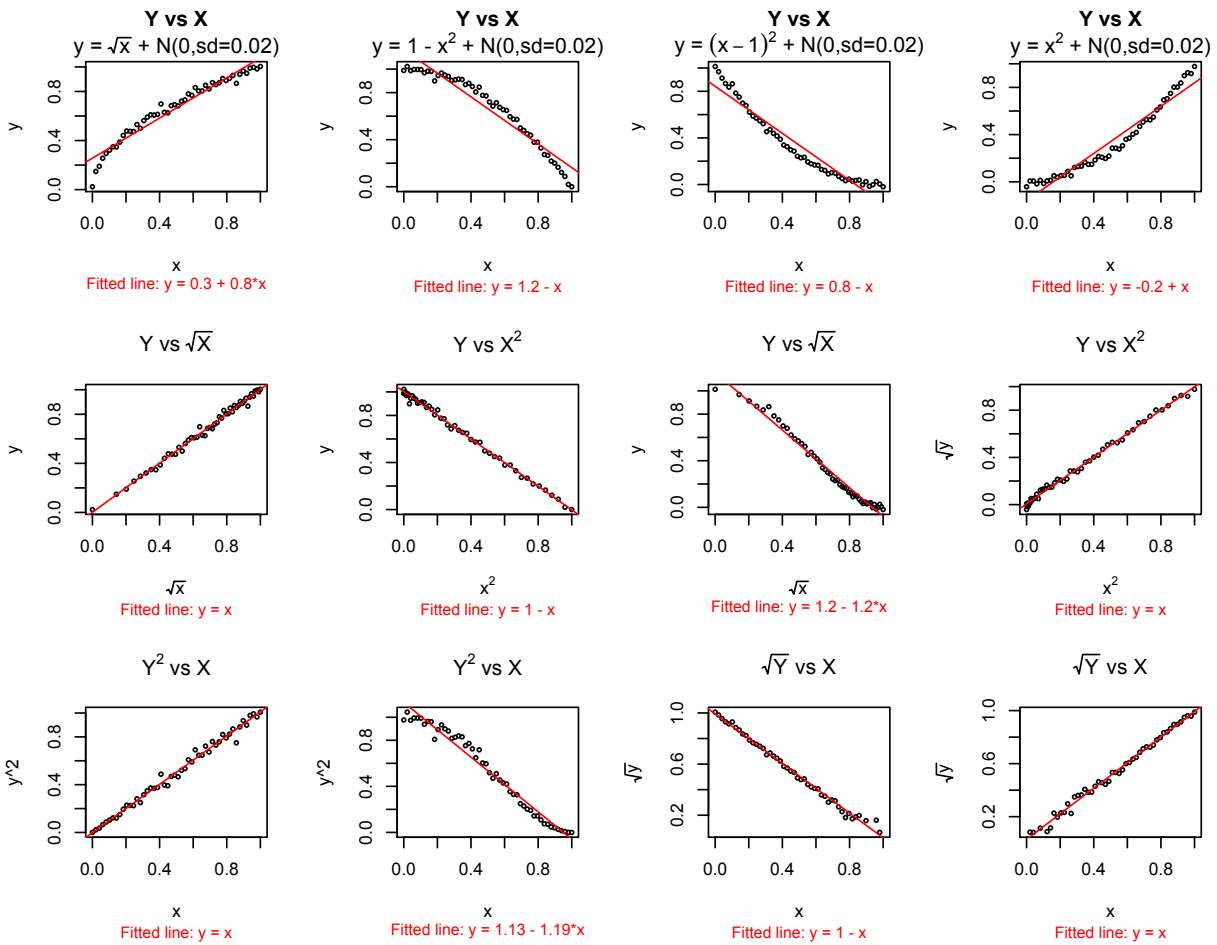
Slope = 1.

Transforming for straightness

Advantage of linear relationship

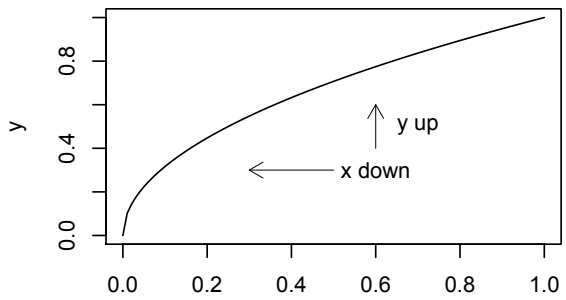
- Easier interpretation
- Departure from fit are more easily detected
- Easy interpolation and extrapolation

Transforming for straightness



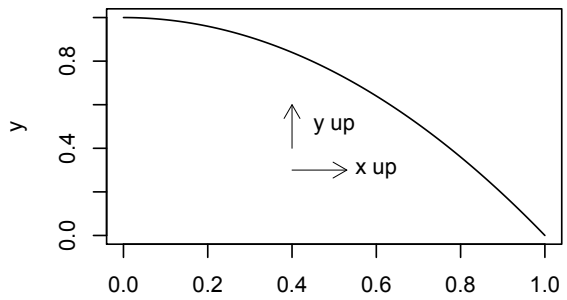
Transforming for straightness

$Y = \sqrt{X}$



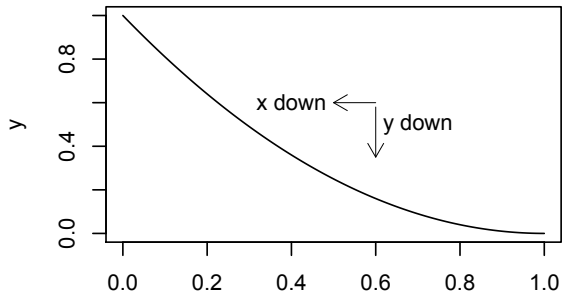
straighten with $\sqrt{x}$ or y^2

$Y = 1 - X^2$



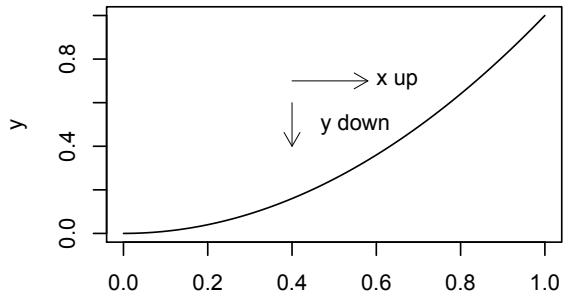
straighten with x^2 or y^2

$Y = (X - 1)^2$



straighten with $\sqrt{x}$ or $\sqrt{y}$

$Y = X^2$



straighten with x^2 or $\sqrt{y}$

Transforming for simple structure (Ch6)

Additive structure for two-way table

Matched transformations

1. Benefits of matched re-expression: first of all, from $T(x)$ to $a + bT(x)$, there presents little additional difficulty in interpretation
    - Can arrange for the re-expressed data to look like the original data, except the extreme values
    - Emphasizes the changes that are due to transformation
    - Helps to compare the effect of different transformations

Matched transformations

2. Procedures for obtaining matched transformation
   - Original data x, nonlinear transformation $y = T(x)$, further transformation $z = a + by = a + bT(x)$.
   - Objective: choose a and b to match z to x.
   - Method 1: choose x_1 and x_2 in original scale, find a and b

$$\begin{cases} z_1 = a + bT(x_1) = x_1 \\ z_2 = a + bT(x_2) = x_2 \end{cases}$$

   - Method 2: choose x_0 (some central value),
     $z_0 = a + bT(x_0) = x_0$

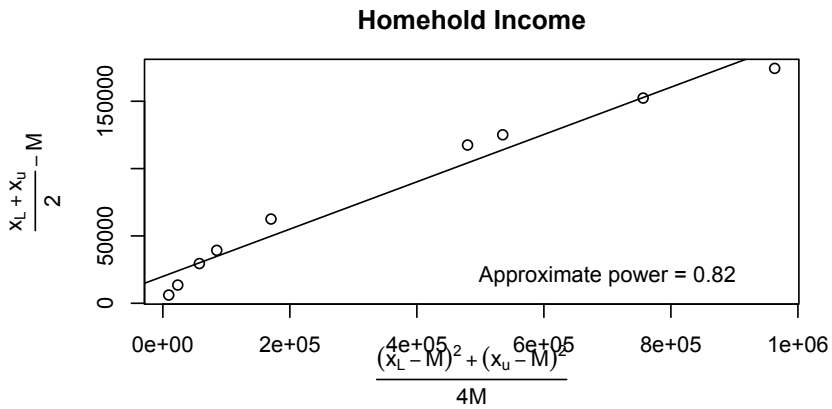
$$\frac{dz}{dx}\Big|_{x_0} = b\frac{dT(x)}{dx}\Big|_{x_0} = 1 \Rightarrow b = \frac{1}{T'(x_0)} \Rightarrow z = x_0 + \frac{T(x) - T(x_0)}{T'(x_0)}$$

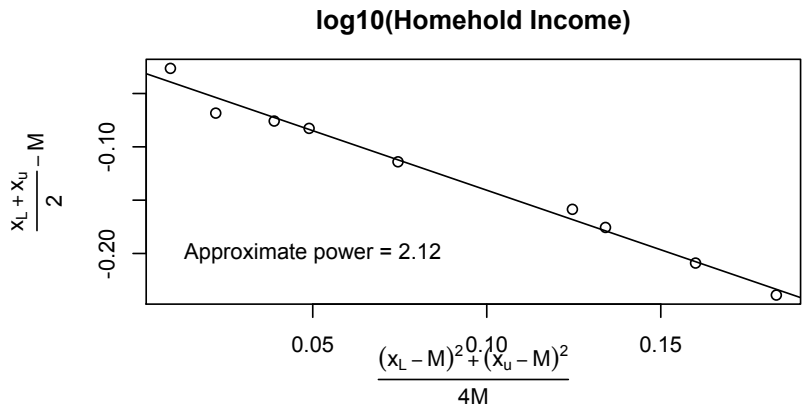
   - Example:
     $T(x) = x^p$, $T'(x_0) = px_0^{p-1}$, $z = x_0 + (x^p - x_0^p)/px_0^{p-1}$
   - Example: $p = 0$, $T(x) = \log_{10}(x)$,
     $T'(x) = (\log_{10}^e / x) = 0.4343/x$,
     $z = x_0 + (\log_{10} x - \log_{10} x_0)x_0/0.4343$

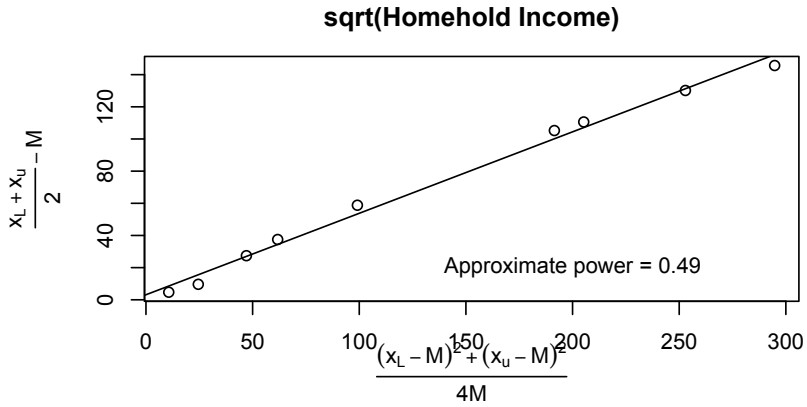
Example: Household Income

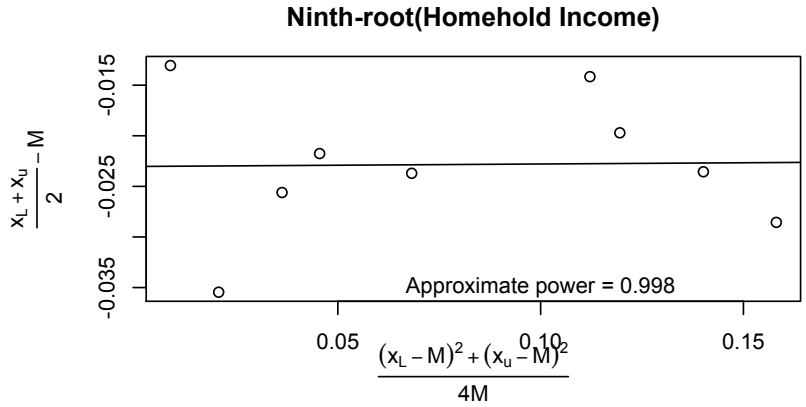
	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	39200	39200	39200	0	0.00
F	98.0	19000	71600	45300	52600	38992.44
E	49.5	12000	93450	52725	81450	35402.29
D	25.0	8400	129000	68700	120600	39305.91
C	13.0	7000	150000	78500	143000	38384.48
B	7.0	4570	198900	101735	194330	45111.72
A	4.0	2380	311000	156690	308620	63828.85
Z	2.5	2100	326500	164300	324400	60975.90
Y	1.5	1600	381500	191550	379900	65826.07
X	1.0	1200	426000	213600	424800	68576.54

Example: Household Income



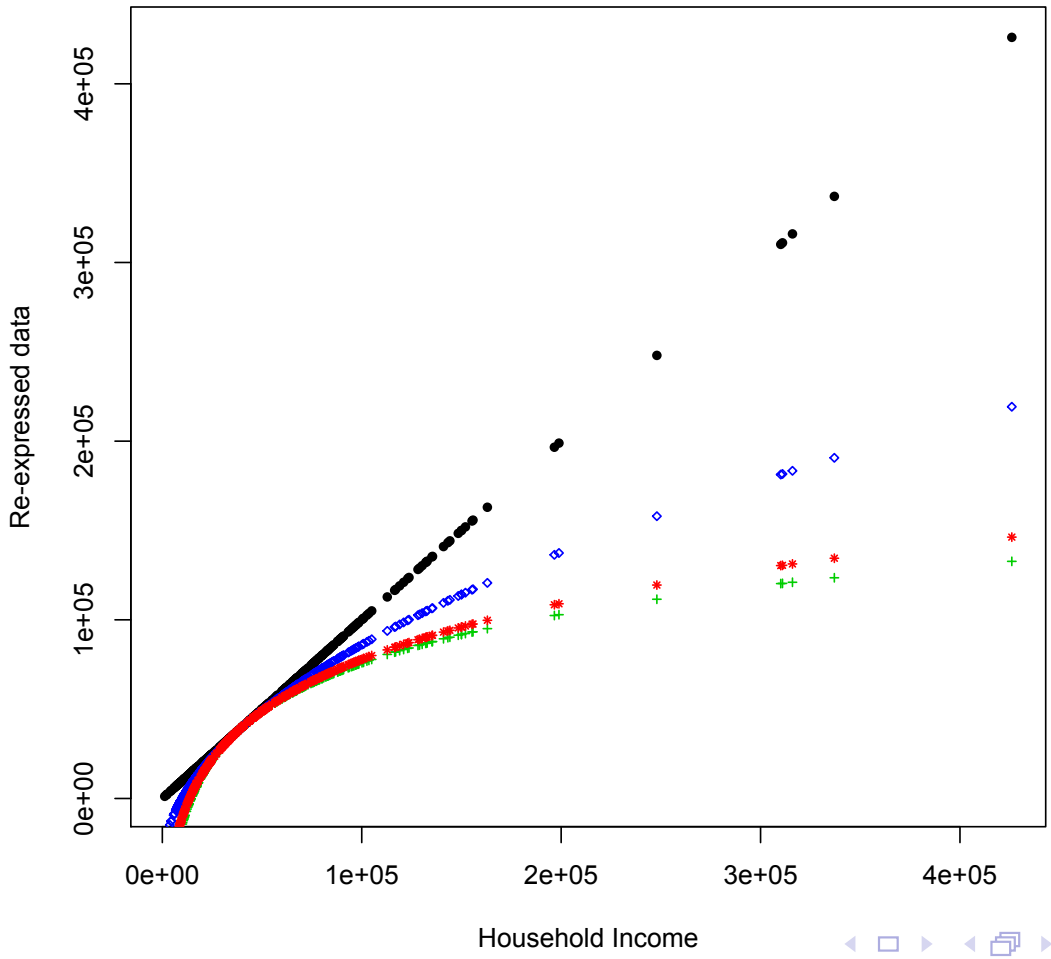






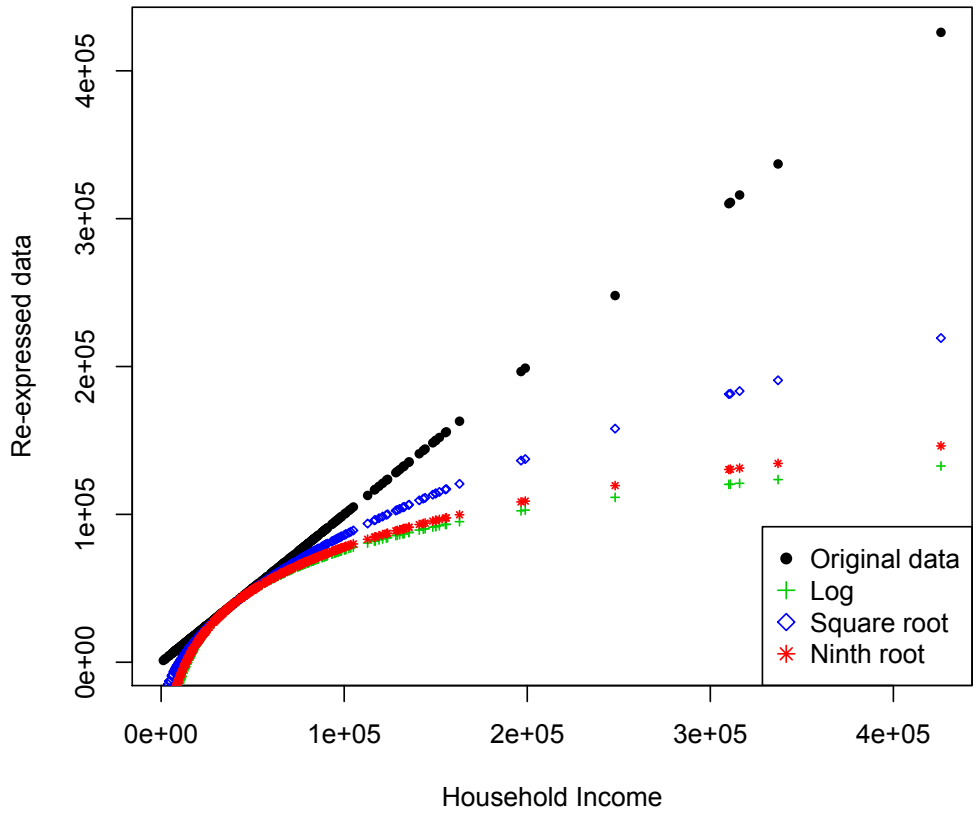
Example: Household Income

Re-expressed household income data versus original data



Example: Household Income

Re-expressed household income data versus original data



Example: Household Income

Matched: log10(Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	39200.000	39200.00	39200.00	0.00	0.00
F	98.0	10810.240	62814.50	36812.37	52004.26	38550.81
E	49.5	-7203.198	73254.42	33025.61	80457.62	34970.95
D	25.0	-21184.678	85891.92	32353.62	107076.60	34898.37
C	13.0	-28331.593	91804.11	31736.26	120135.70	32247.18
B	7.0	-45046.140	102864.91	28909.38	147911.05	34336.04
A	4.0	-70620.394	120386.72	24883.16	191007.12	39504.13
Z	2.5	-75571.221	122272.99	23350.88	197844.21	37187.82
Y	1.5	-87451.332	128127.38	20338.02	215578.71	37353.77
X	1.0	-97463.387	132720.71	17628.66	230184.09	37159.20

Example: Household Income

Matched: Square-root (Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	39200.000	39200.00	39200.00	0.00	0.00
F	98.0	15382.048	66756.97	41069.51	51374.92	38084.29
E	49.5	4177.413	81849.30	43013.36	77671.89	33760.13
D	25.0	-2907.852	103022.36	50057.25	105930.21	34524.74
C	13.0	-6069.953	114162.32	54046.18	120232.27	32273.10
B	7.0	-12431.063	137399.89	62484.41	149830.95	34781.72
A	4.0	-19882.029	181627.53	80872.75	201509.56	41676.25
Z	2.5	-21059.073	187034.30	82987.61	208093.38	39114.30
Y	1.5	-23487.051	204961.89	90737.42	228448.94	39583.83
X	1.0	-25482.857	219250.77	96883.96	244733.63	39507.97

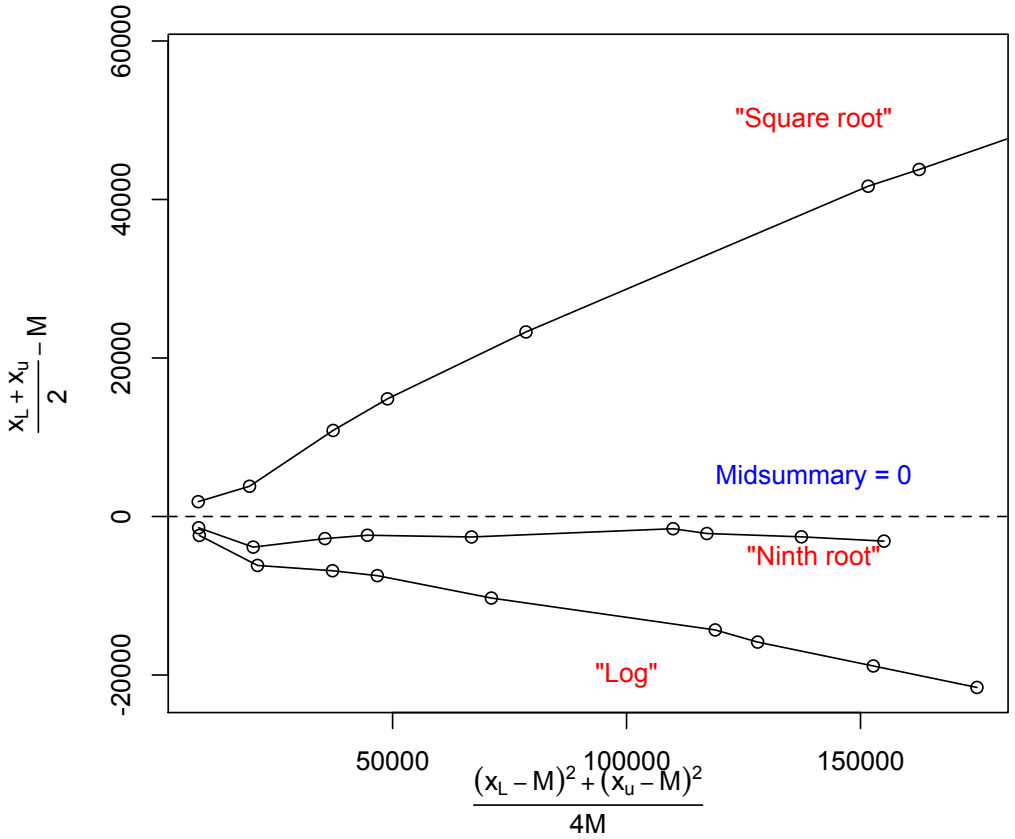
Example: Household Income

Matched: Ninth-root (Household Income)

	Depth	Lower	Upper	Mid	Spread	pseudo-s
M	195.0	39200.000	39200.00	39200.00	0.00	0.00
F	98.0	11922.135	63623.06	37772.60	51700.93	38325.96
E	49.5	-4281.555	74952.67	35335.56	79234.22	34439.20
D	25.0	-16300.287	89123.30	36411.51	105423.59	34359.62
C	13.0	-22262.376	95929.05	36833.33	118191.42	31725.29
B	7.0	-35743.338	108971.94	36614.30	144715.28	33594.17
A	4.0	-55172.530	130489.16	37658.31	185661.69	38398.58
Z	2.5	-58770.178	132872.75	37051.28	191642.92	36022.19
Y	1.5	-67112.727	140379.03	36633.15	207491.76	35952.53
X	1.0	-74105.962	146289.41	36091.73	220395.38	35578.98

Example: Household Income

Matched midsummaries for the household income data:



Matched transformations

3. Matching expected values

$$\frac{\sum T(X_i)}{n} \neq T(\bar{x}), \quad E(T(x)) \neq T(EX)$$

Matching can reduce the difference

$$\begin{aligned} E(Z) &= a + bET(X) \\ &\approx x_0 - \frac{T(x_0)}{T'(x_0)} + \frac{1}{T'(x_0)} \\ &\qquad \cdot \left[T(x_0) + E(X - x_0)T'(x_0) + \frac{1}{2}E(X - x_0)^2 T' \right. \\ &= E(X) + \frac{1}{2}E(X - x_0)^2 \frac{T''(x_0)}{T'(x_0)} \end{aligned}$$

Matched transformations – Interpretation:

We can write
$E(X - x_0)^2 = E(X - EX)^2 + (EX - x_0)^2 = var(X) + (EX - x_0)^2$.

If choose x_0 to be near the mean of X

$$E(Z) \approx E(X) + \frac{1}{2} var(X) \frac{T''(x_0)}{T'(x_0)}$$

If Z is a transformation of X matched at a point near the mean of X, the mean of Z is approximately equal to the mean of X plus a term that depends on the spread of X and the curvature of the transformation.

What we have learned about transformation?

- Transformation for one batch – symmetry
- Transformation for multiple batches – equal spread
- Transformation for y vs x – straightness

Serendipitous effects of transformation

1. Data that are amounts or counts usually display both increasing spread with increasing level and right-skewness
   - Bounded below by zero: variation of spread with level; right-skewness
   - Poisson variable, mean λ, variance λ
2. Transformation for stable spread will necessarily compress the scale more for larger values than smaller values
3. Transformation for symmetry will also compress the scale more for larger values than for smaller values
4. Thus transformation for either spread or symmetry will usually help us towards both objective simulataneously.

When is transformation worthwhile?

1. Range consideration: Amounts or counts, when the range of batch is relatively large.

   Rule of thumb: transformation helpful when the ratio

   $$\frac{\text{largest data value}}{\text{smallest data value}} > 20$$

   not helpful if the ratio is smaller than 2
2. Looking at residuals (y-versus-x in Ch5, two-way table in Ch6)
3. Using transformation plots: slope ≈ 0, no transformation; if plot nearly linear & slope $\neq 0$, power; if nonlinear, complex transformation.
   - spread vs level plots
   - transformation plot for symmetry
   - additional one in Ch6
4. Trial and error

Data and transformation

Types of data:

1. Counts
2. Amounts (quantitative)
3. Balances (differences, ratios, log(ratios))
4. Counted fractions and percentages
5. Categories $\rightarrow$ Ranks

Possible re-expressions:

1. Counts: Square roots
2. Amounts: nothing if nearly symmetric; else Ladder
3. Balances: often none; but re-expression of amounts or counts before subtraction sometimes helps
4. Counted fractions and percentages: re-expression is often helpful, special techniques
5. Ranks: Normal scores $\Phi^{-1}(p_i)$, $p_i = \dfrac{i - 1/3}{n + 1/3}$

Transformations for fractions ($0 \leq p \leq 1$) (EDA Ch15)

- Counted fraction: $p =$ (# successes) / (total # observations)
- Objective: compare, analyze, describe distributions (sequence of counted fractions)
- Cutting value or cut: a value that divides "below" from "above"
- Instead of (count below cut-off)/n, use

$$(\text{count below} + \frac{1}{2}\text{count equal} + \frac{1}{6})/(\text{total count} + 1/3)$$

- Example: $n = 6$, $\{X\} = \{0, 1, 2, 2, 3, 7\}$
  1. cut $= 2$, $P\{x \leq 2\} = (2 + 1/2 \times 2 + 1/6)/(6 + 1/3) = 0.5$
  2. cut $= 3$, $P\{x \leq 3\} = (4 + 1/2 \times 1 + 1/6)/(6 + 1/3) = 0.737$
  3. cut $= 4$, $P\{x \leq 4\} = (5 + 1/6)/(6 + 1/3) = 0.816$

Three matched scales

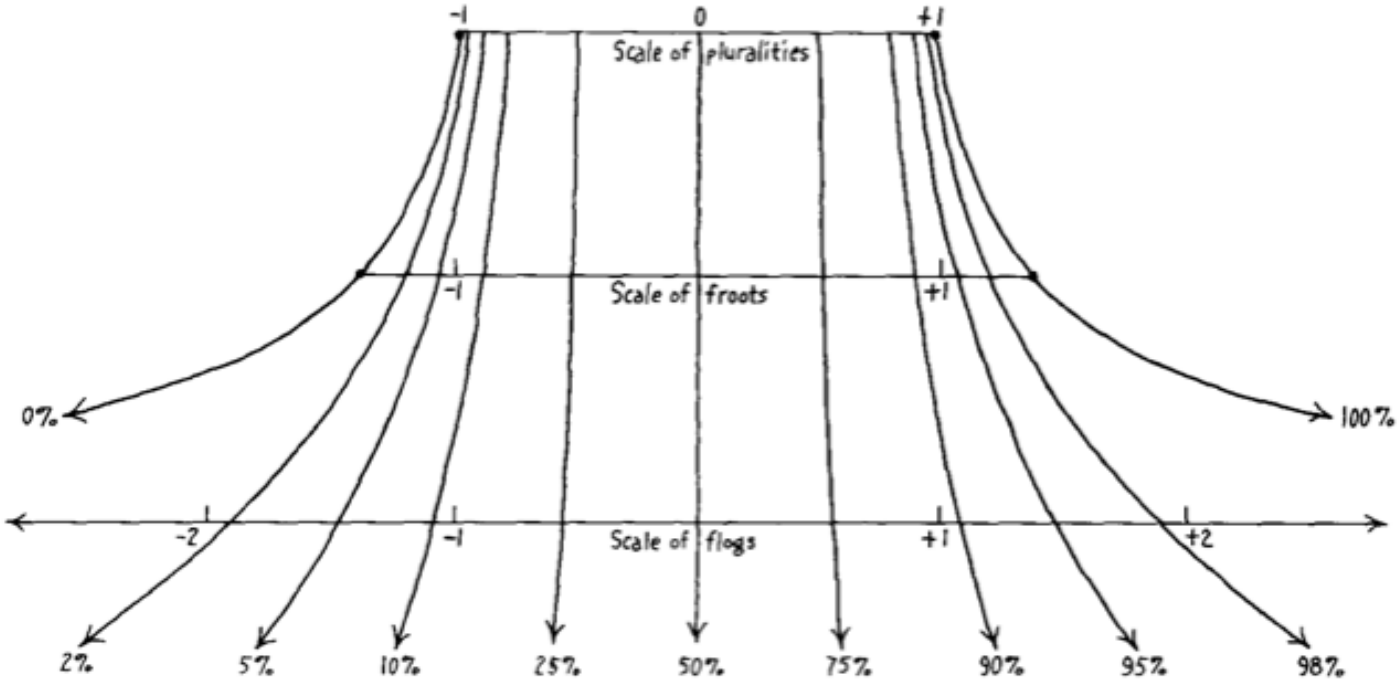
Suppose f is the fraction concerned, we want new scales to have nice properties:

- $50\% \rightarrow 0$
- swap of f with $1 - f \rightarrow$ change the sign, not the size

Matched transformations for f:

- plurality: $f - (1 - f)$
  Or cents (folded percents): (% yest) - (% no)
- froots (folded roots): $\sqrt{2f} - \sqrt{2(1 - f)}$
- flogs (folded logs): $(1/2)\ln(f) - (1/2)\ln(1 - f)$
- Both (froots and flogs) stretch ends of scales

Illustration



Example: comparison within one dataset

Example: Gallup Poll results (Washington Post, Feb 62)

Protestants shift support to Kennedy

	Protestants		Catholics	
Date	Nov 1960	Jan 1962	Nov 1960	Jan 1962
Kennedy	38%	59%	78%	89%
Nixon	62%	41%	22%	11%

"One of the major reasons for the President's popularity has been his success in allaying anti-Catholic sentiment while not losing the support of fellow Catholitcs" (Protestants: 21% increase; Catholics: 11% increase).

Example

What is really going on?

- Scale stretching near the end (38% $\rightarrow$ 42%, 89% $\rightarrow$ 93%)
- Take flogs:

Data	Protestant	Catholics	Difference
Nov' 60	-0.24	+0.63	+0.87
Jan' 62	+0.18	+1.05	+0.87
Change	+0.42	0.42	

- JWT: "JFK's popularity increased by 0.42 on the flog scale, regardless of whether you asked Protestants or Catholics".
- flog: not this good all the time, but helps often

HW 3 will be posted soon.