Exploratory Data Analysis

Lecture 2: Letter Values & Letter Value Plots

David B King, Ph.D.

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

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

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

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

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

  - range

Letter values as measures

- Location summary: median, trimean

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

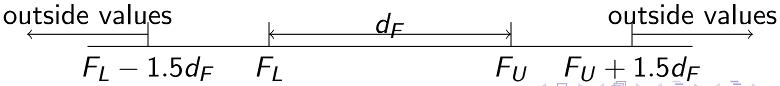
  $$\text{trimean} = \tfrac{1}{4}(\text{lower fourth}) + \tfrac{1}{2}(\text{median}) + \tfrac{1}{4}(\text{upper fourth})$$

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

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

  - 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

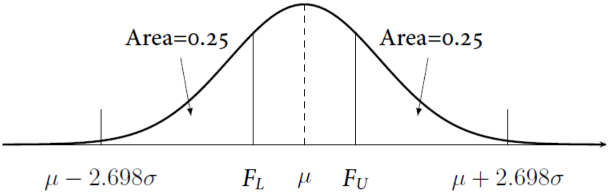
  Rule of thumb:



Letter values: Gaussian distribution

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:

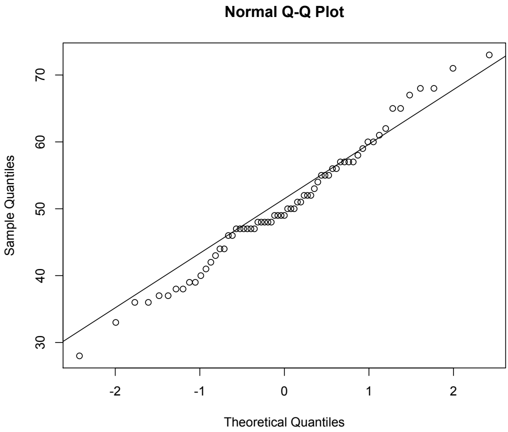
```
> stem(data1)

  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:

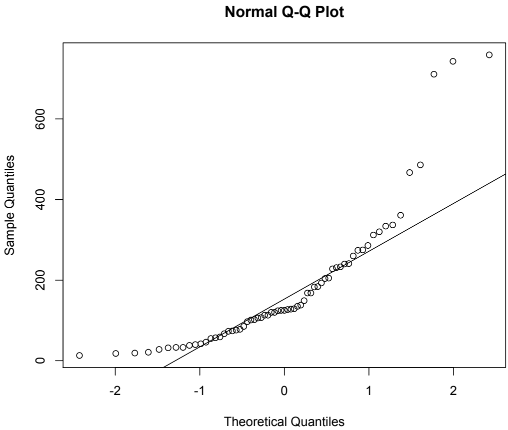
```
> stem(data2)

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

  0 | 1222333444566677789
  1 | 00011112222333344577889
  2 | 0133344 6789
  3 | 12346
  4 | 79
  5 |
  6 |
  7 | 146
```

Letter values: display

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



Summary and more theoretical results about letter values

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,

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

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

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

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

  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

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

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 $n \to \infty$

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

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

2. If we take $q = (1/2)^r$ and $p = (1/2)^{(r+1)}$ then $\mathrm{Corr}^2\left[X_{(i)}, X_{(j)}\right] = \frac{(1/2)^{r+1}}{(1-(1/2)^{r+1})}\frac{(1-(1/2)^r)}{(1/2)^r} \to 1/2$ 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.

- Next lecture: Boxplot, letter-value boxplot
- Homework 2 Assignment coming up.

R code for letter value display

- Back-to-back stem-and-leaf
- lval()
- lval.sub()