PSQF 4143: Section 2

Brandon LeBeau

Summarizing Data

- Data in its raw form is often too complex to summarize and understand quickly.
- As such, summarizing the data with descriptive statistics (section 3) or with tables and graphs (this section) can be very helpful.

Motivating Example

```
##                    title length
## 1          13 Going On 30     98
## 2          50 First Dates     99
## 3              Anchorman    104
## 4           Aviator, The    170
## 5       Butterfly Effect    120
## 6    Cinderella Story, A     95
## 7             Collateral    120
## 8                  Crash    113
## 9       Dawn of the Dead    109
## 10             Dodgeball     92
## 11      Eternal Sunshine    108
## 12        Girl Next Door    110
## 13 Harry Potter: Azkaban    141
## 14               Hellboy    132
## 15      Incredibles, The    121
```

```
##                          title length
## 1          Kill Bill: Vol. 2    136
## 2                King Arthur    140
## 3                 Mean Girls     97
## 4        Million Dollar Baby    132
## 5          Napoleon Dynamite     86
## 6              Notebook, The    123
## 7  Phantom of the Opera, The    143
## 8              Punisher, The    124
## 9                        Saw    100
## 10         Shaun of the Dead     99
## 11             Spider-Man 2    127
## 12                     Troy    162
## 13              Van Helsing    132
## 14             Village, The    108
## 15             White Chicks    109
```

Ungrouped Frequency Table

Ungrouped Frequency Table Creation Steps

- To create an ungrouped frequency table:
    1. List all numbers from the minimum to maximum
    2. Count the number that fall within each number
    3. Numbers can have 0 frequency
- Strengths:
    - Able to see all the numbers in the distribution
    - Can quickly glance at the range
    - Can quickly see which number is most frequent.
- Weaknesses:
    - Tables can be large depending on range
    - Lose some information, namely who had which value

Grouped Frequency Table

Grouped Frequency Table Creation Steps

- To create a grouped frequency table:
  1. First need to find the class interval size

$$i = RealUpperLimit - RealLowerLimit$$

  2. List out all the class intervals
  3. Count the number that fall within each interval
- Strengths:
  - Can quickly see where most values fall
  -
- Weaknesses:
  - No longer know exact values
  - The choice of class interval size can influence the table

Effect of Class Interval Size

Var1	Freq
[0,5)	34
[5,10)	98
[10,15)	111
[15,20)	84
[20,25)	57
[25,30)	47
[30,35)	23
[35,40)	9
[40,45)	15
[45,50)	7
[50,55)	4
[55,60)	12
[60,65)	16
[65,70)	8
[70,75)	35
[75,80)	44
[80,85)	116
[85,90)	184
[90,95)	280
[95,100)	215
[100,105)	146
[105,110)	104
[110,115)	62
[115,120)	49
[120,125)	55
[125,130)	34
[130,135)	13
[135,140)	17
[140,145)	17
[145,150)	5
[150,155)	8
[155,160)	4
[160,165)	7
[165,170)	0
[170,175)	6
[175,180)	4
[180,185)	2
[185,190)	2
[190,195)	2
[195,200)	0

Var1	Freq
[0,10)	132
[10,20)	195
[20,30)	104
[30,40)	32
[40,50)	22
[50,60)	16
[60,70)	24
[70,80)	79
[80,90)	300
[90,100)	495
[100,110)	250
[110,120)	111
[120,130)	89
[130,140)	30
[140,150)	22
[150,160)	12
[160,170)	7
[170,180)	10
[180,190)	4
[190,200)	2

Var1	Freq
[0,20)	327
[20,40)	136
[40,60)	38
[60,80)	103
[80,100)	795
[100,120)	361
[120,140)	119
[140,160)	34
[160,180)	17
[180,200)	6

Var1	Freq
[0,50)	485
[50,100)	914
[100,150)	502
[150,200)	35

Additional Notes for Grouped Frequency Tables

- The class intervals should be mutually exclusive
- For quantitative values, no gaps between class intervals
- All class intervals should be the same size
- On average, 10 to 20 class intervals are most appropriate
- Common class interval sizes include: 1, 2, 3, 5, 10, 15, 20, 25, ...
- The lower limit should equal the class size times a constant integer (i.e. 5 * 10 = 50)

Unequal Frequency Distributions

- One area where unequal frequency distributions can be useful is with variables such as income.
- Income tends to have more individuals at the low end and fewer at the high end.
- As such, it can be helpful to have larger class intervals at the upper end of the income scale.
- Unequal intervals can be used for tables, but not for graphs.

Unequal Frequency Distribution Example

Table 3.8 Frequency Distribution 1,000 Individual Incomes

Annual Income	*f*	Class Size
50,000–99,999	1	50,000
25,000–49,999	2	25,000
20,000–24,999	2	5,000
15,000–19,999	4	5,000
10,000–14,999	5	5,000
7,000– 9,999	6	3,000
5,000– 6,999	8	2,000
4,000– 4,999	14	1,000
3,500– 3,999	17	1,000
3,000– 3,499	41	500
2,500– 2,999	85	500
2,000– 2,499	116	500
1,500– 1,999	124	500
1,250– 1,499	75	250
1,000– 1,249	78	250
750– 999	99	250
500– 749	104	250
250– 499	107	250
0– 249	112	250
	1,000	

Figure 1: Unequal Income

Relative Frequency Distributions

- These are frequency distributions with proportion or percentage of each class interval
- Useful for comparisons of two or more groups especially when the number in each group differ

Relative Frequency Distribution Example

Relative Frequency by Groups

```
##
##                PG   R
##     [10, 25)    0   1
##     [25, 40)    0   1
##     [40, 55)    0   0
##     [55, 70)    1   0
##     [70, 85)    3  10
##    [85, 100)   23 111
##   [100, 115)    7  40
##   [115, 130)    5  21
##   [130, 145)    3   8
##   [145, 160)    0   2
##   [160, 175)    0   1
```

Cumulative Frequency Distributions

Cumulative Frequency Creation

1. Start with a basic frequency distribution, grouped or ungrouped.
2. Starting with the smallest value, keep a running tally of the number encountered.
    - This can be done by taking the frequency for the current category plus all prior categories
3. Optional - add columns for cumulative proportion or cumulative percentage.

Frequency Distribution for Qualitative Variables

```
##
##       NC-17    PG PG-13     R
## 53864    16   528  1003  3377
```

Graphs

- Graphs are a great alternative to many of the tables we discussed above as they tend to be easier to quickly interpret and understand.
- One benefit of graphs is the ability to explore the shape of distributions.
- However, it is also easier to create misleading graphics.
- Graphs for quantiative variables:
    - Histograms
    - Frequency Polygons
    - Cumulative Polygon (Ogive)
    - Stem and Leaf
- Graphs for qualitative variables:

  - Bar Graphs
  - Pie Charts
- Note: there are many other graphs that can be used too that we are not discussing.

Histograms

- A histogram is a visual representation of a frequency table for quantitative variables.
- No gaps between bars as the x-axis is continuous.

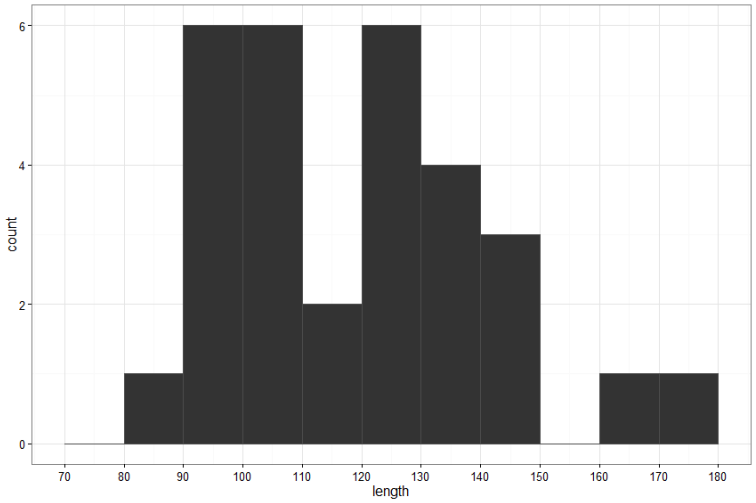


Figure 2: plot of chunk hist

Shapes of Distributions

- Distributions can be categorized based on their symmetry (skewness) and kurtosis.
- Symmetry/Skewness is the amount a distribution leans one way or the other.
  - A symmetric distribution is one where each half of the distribution are mirror images of one another.
  - An asymmetric (skewed) distribution is one where they are not mirror images.
  - A positively (right) skewed distribution is one where the bulk of data lie in the lower portion with a long upper tail.
  - A negatively (left) skewed distribution is one where the bulk of data lie in the upper portion with a long lower tail.
- Kurtosis refers to the peakedness of the distribution. Relatedly, this also refers to the portion of the distribution that resides in the tails.
  - Mesokurtic refers to an intermediate distribution with average tails and peakedness.

  - Platykurtic refers to a distribution that is flatter with more observations in the tail.
  - Leptokurtic refers to a distribution that is steeper with fewer observations in the tail.

Describing Distributions

- Location (Central Tendency)
  - Where is the middle score?
  - Where are the scores concentrated?
- Variability
  - Dispersion
  - Spread
  - Range

Examples of common shapes

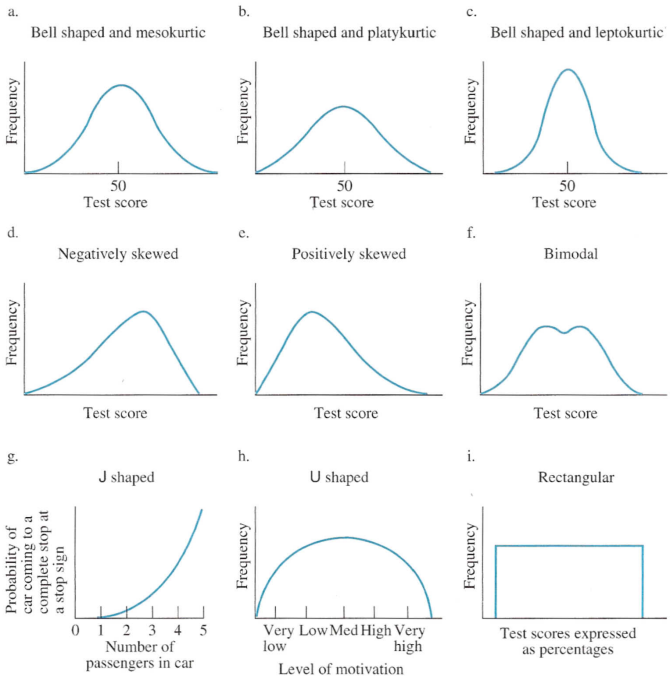
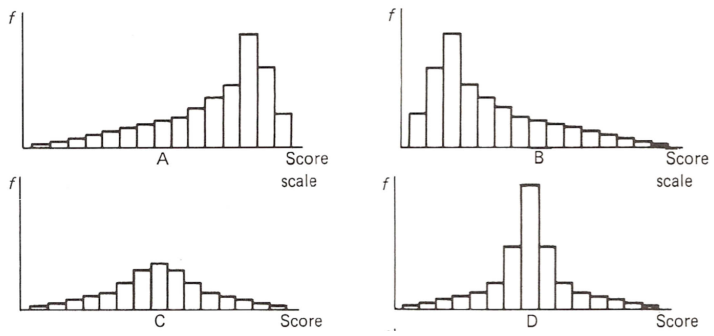


Figure 2.6-1. Common distributions in behavioral and educational research.

Figure 3: Distribution Shapes

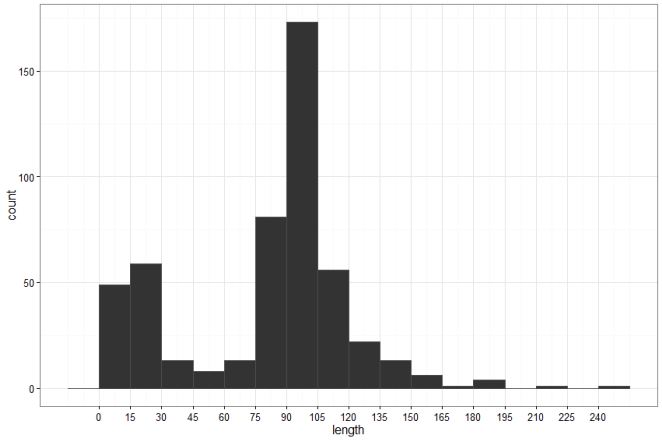
Shapes with Histograms

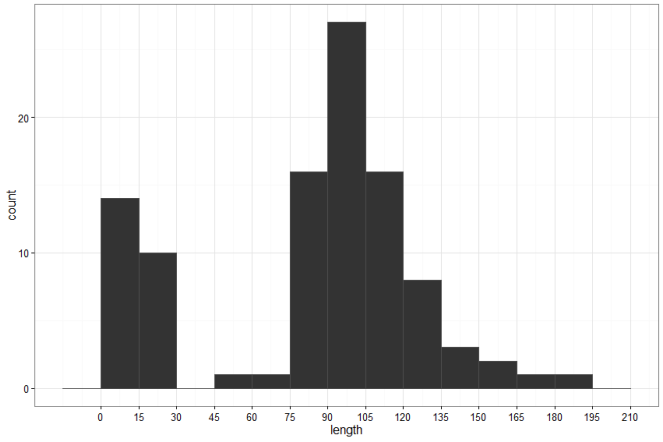


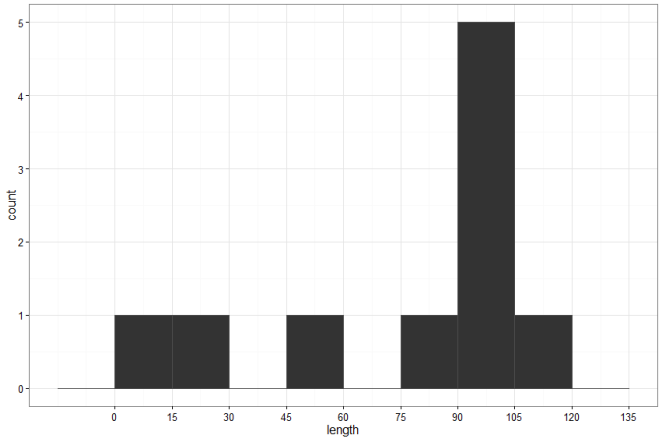
Factors Affecting Distribution Shape

1. Sampling
    - The larger the sample size, the closer it will approximate the population.
2. Relative Scale
    - The height of the histogram should be 3/4 of the width.
3. Frequency Scale
    - Always continuous and start at 0.
4. Interval size / number of intervals

Effect of Sampling







Effect of Relative Scale

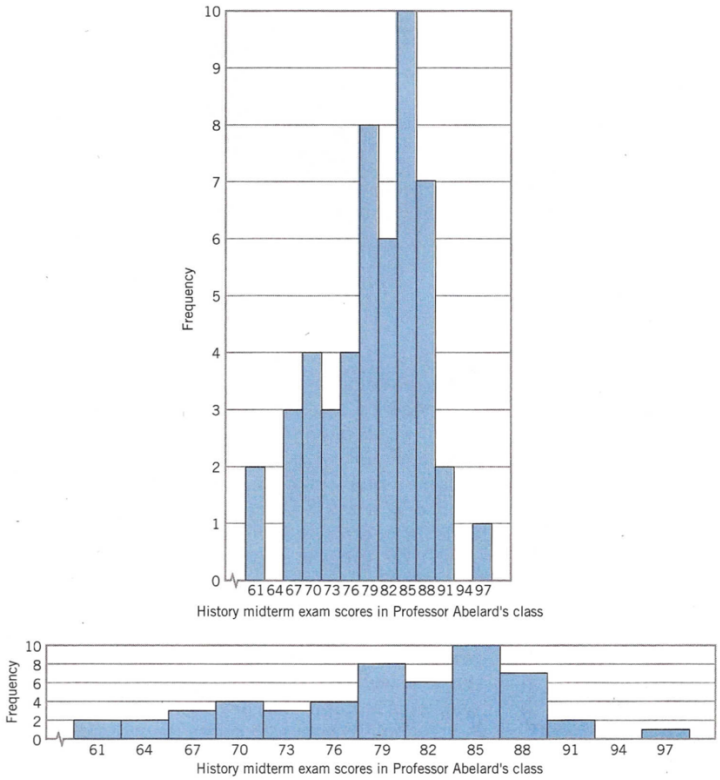
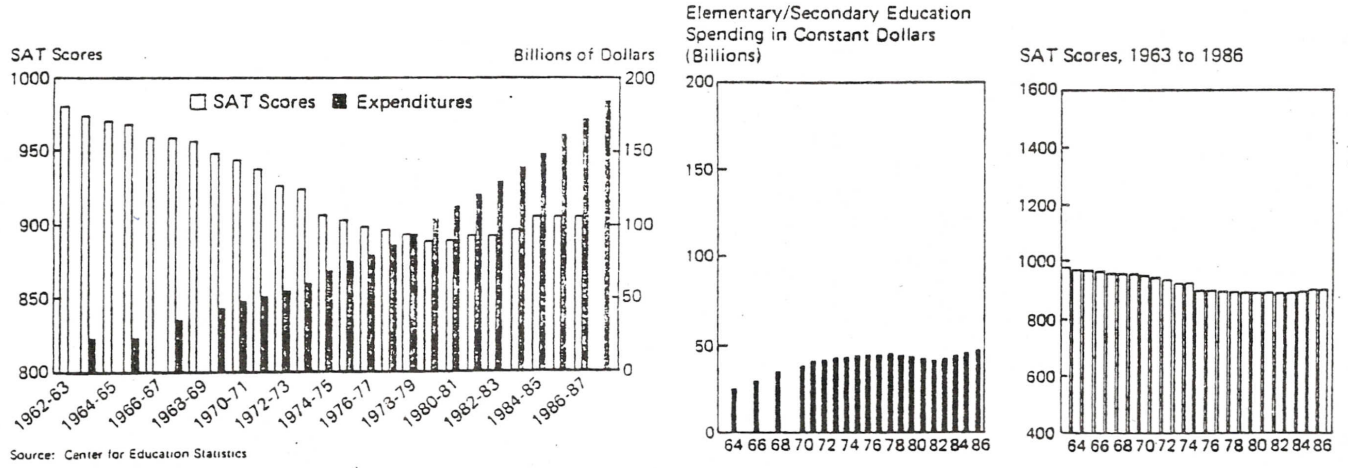
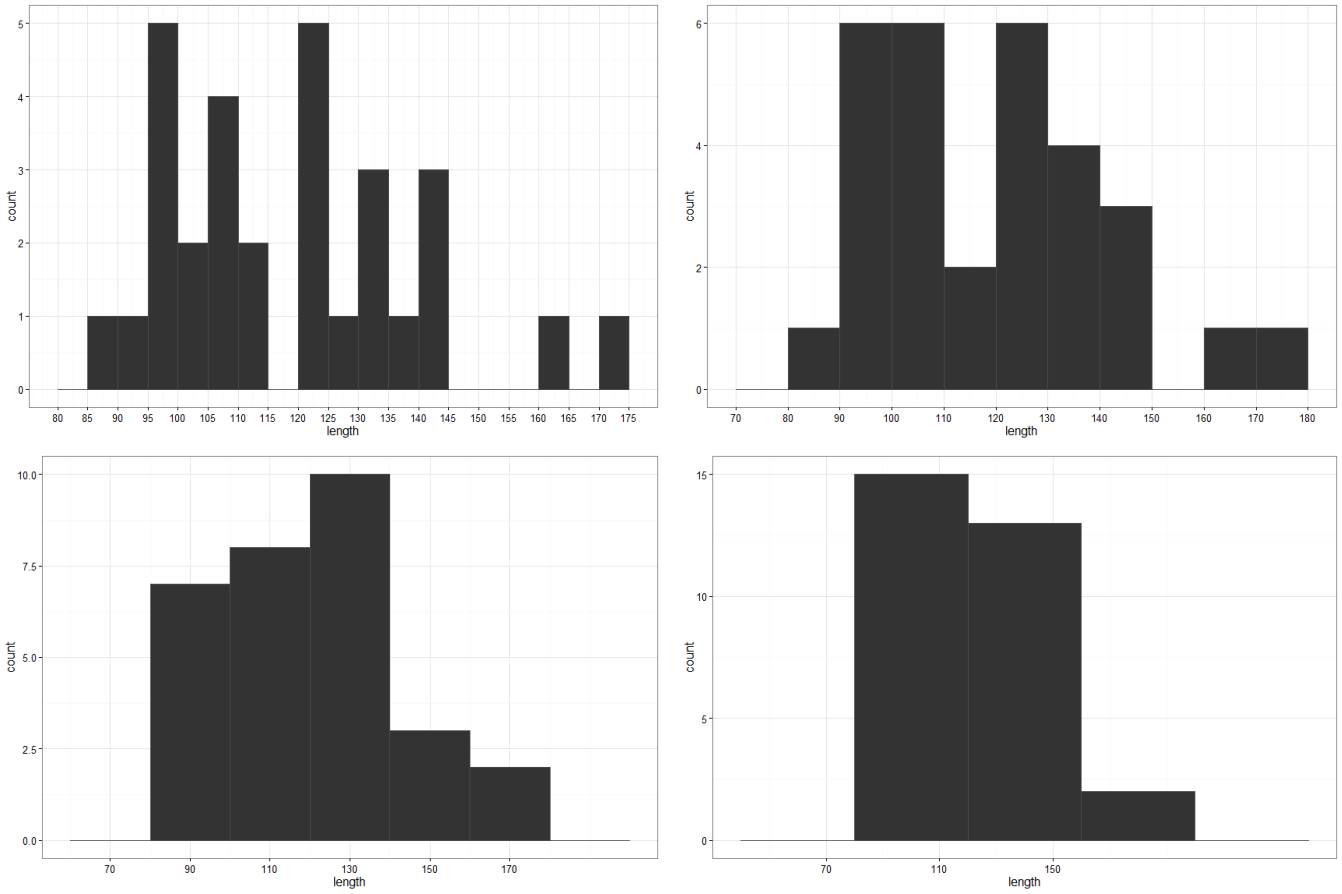


Figure 4: Relative Scale

Effect of Frequency Not Starting at Zero



Effect of Binwidth on Histograms



Cumulative Polygon (Ogive)

- The cumulative polygon or ogive, is a graphical representation of the cumulative frequency.
- This plot is unique in that it never decreases.

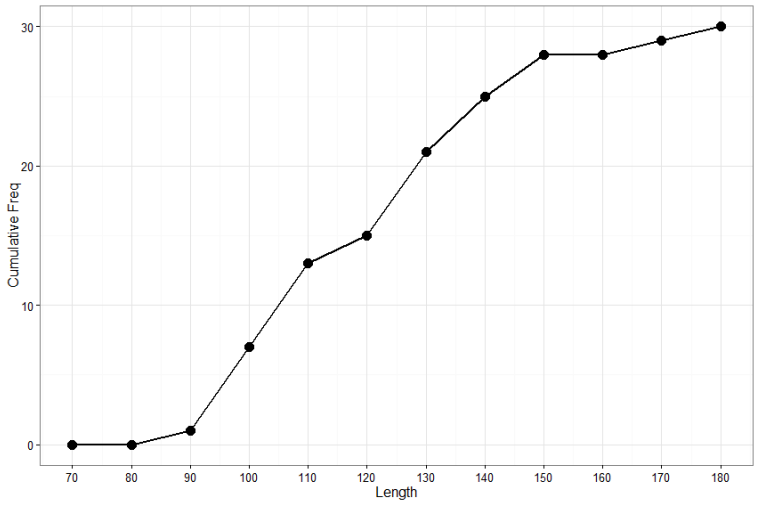


Figure 5: plot of chunk ogive

Ogive - Cumulative Percentage

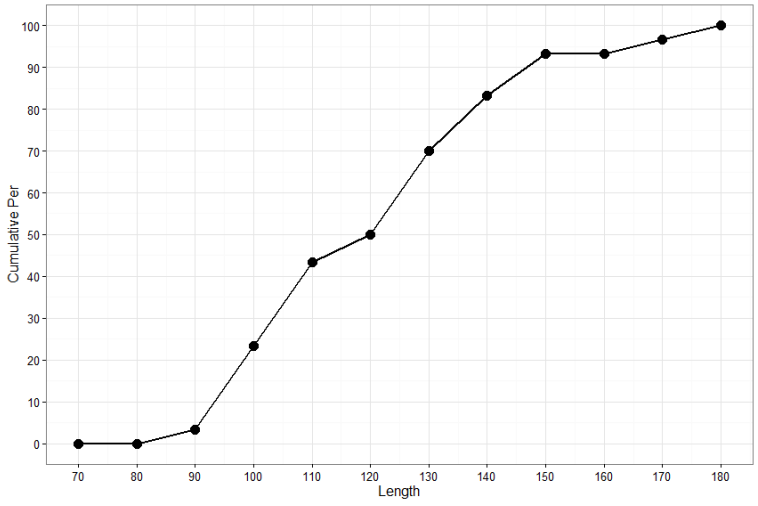


Figure 6: plot of chunk ogivepercent

Real Life Ogive Example

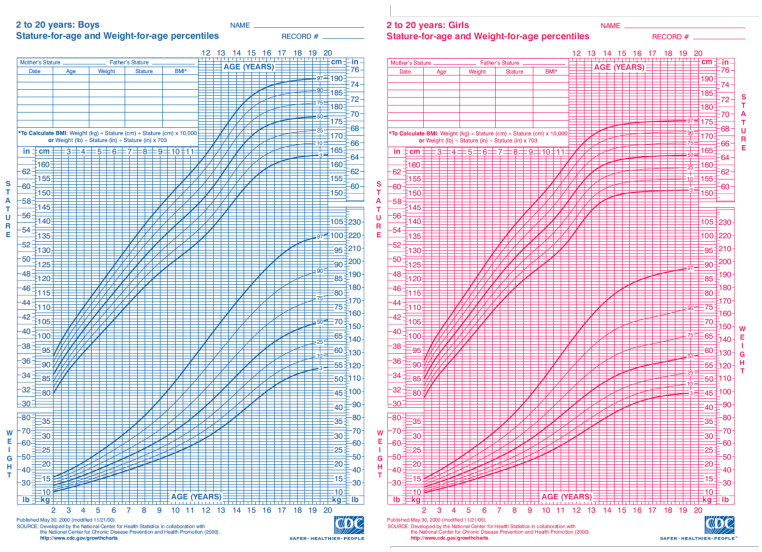


Figure 7: Ogive Examp

Percentile Ranks and Percentiles

- Percentile Ranks:
  - The percentage of scores at or below a given point.
  - Denoted by $P_R(X)$, read as the percentile rank of score X.
- Percentiles:
  - The point on the score scale below which a specified percentage of scores fall.
  - Denoted by $P_{\%}\%$
- A percentile is the inverse of a percentile rank.
- The percentile rank of a given point on the score scale is the percentage of scores falling below this point in the ordered series of scores.
  - The value of the point iteself is the percentile corresponding to this percentile rank.
  - Example: if $P_R(X) = 60$, then $P_{60} = X$.
- Never say, 'My score was **in** the top quartile'.
  - A quartile (or percentile or decile) is a point on the score scale, not an interval.
  - Instead say, 'My score was at the top quartile'.

Special Percentile Points

- Deciles:
  - $D_1 = P_{10}$ - First decile
  - $D_2 = P_{20}$ - Second decile
  - $D_5 = Q_2 = P_{50} = Mdn$ - Median
  - $Q_1 = P_{25}$ - First Quartile
  - $Q_3 = P_{75}$ - Third Quartile

Estimate Percentile Ranks/Percentiles

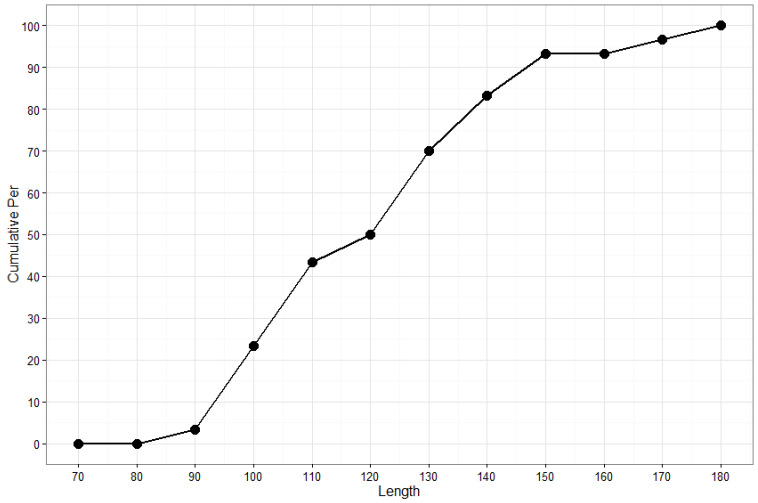


Figure 8: plot of chunk ogivepercent2

- $P_R(122) =$
- $P_R(160) =$
- $P_{50} =$
- $P_{80} =$

Decile Differences

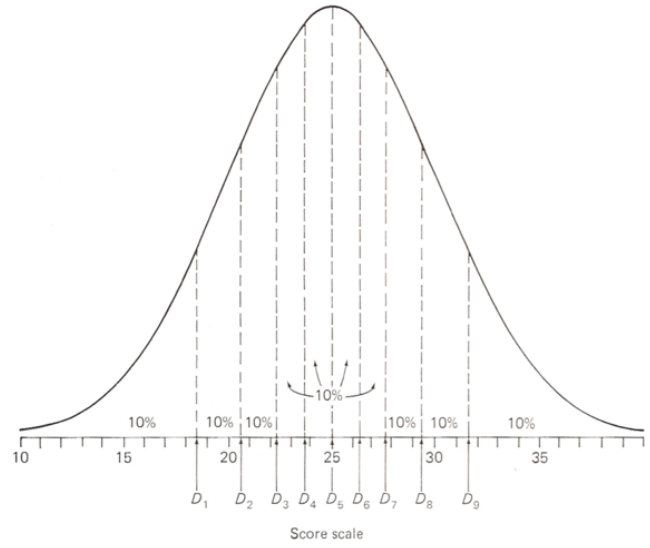


Figure 4.5 Smoothed polygon of idealized population (unimodal symmetrical) of measures of reading comprehension showing locations of nine decile points

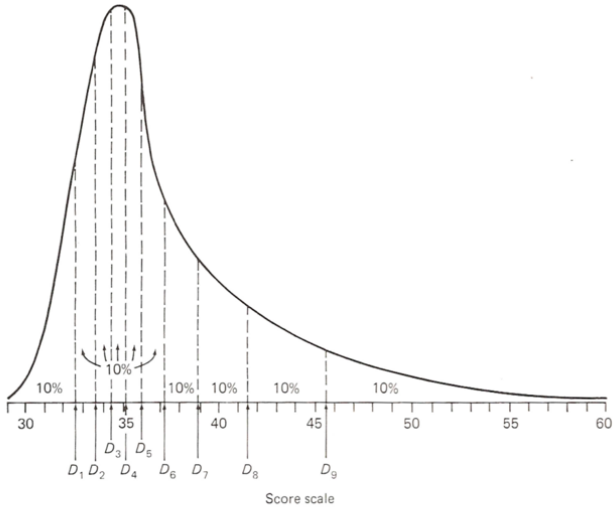


Figure 4.6 Smoothed polygon of idealized population (positively skewed) of measures of reading comprehension showing locations of nine decile points

Table 4.4 Decile Points and Interdecile Distances (Unimodal and Symmetrical Distribution)

Decile	Point	Distance between Points
D_9	31.5	
		2.2
D_8	29.3	
		1.6
D_7	27.7	
		1.4
D_6	26.3	
		1.3
D_5	25.0	
		1.3
D_4	23.7	
		1.4
D_3	22.3	
		1.6
D_2	20.7	
		2.2
D_1	18.5	

Table 4.5 Decile Points and Interdecile Distances (Skewed Distribution)

Decile	Point	Distance between Points
D_9	45.6	
		4.2
D_8	41.4	
		2.5
D_7	38.9	
		1.7
D_6	37.2	
		1.2
D_5	36.0	
		0.9
D_4	35.1	
		0.7
D_3	34.4	
		0.8
D_2	33.6	
		1.0
D_1	32.6	

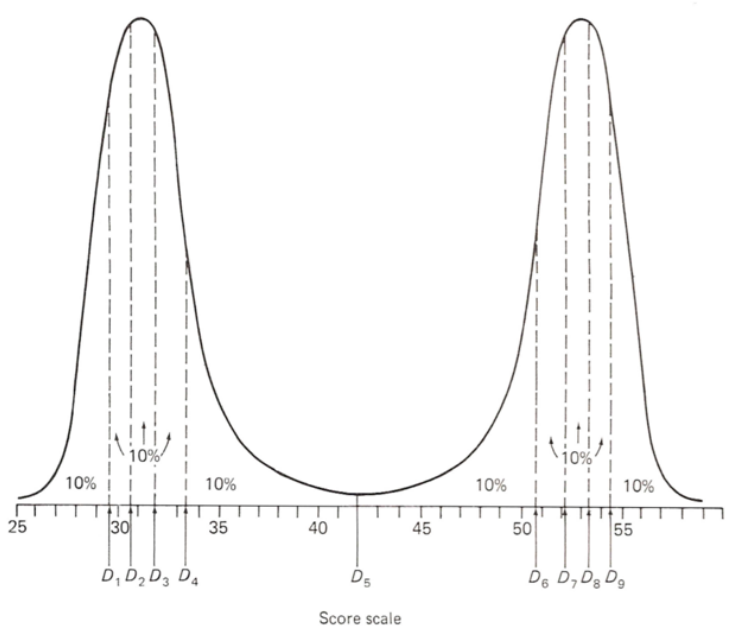


Figure 4.7 Smoothed polygon of idealized population (Bimodal) of measures of reading comprehension showing locations of nine decile points

Table 4.6 Decile Points and Interdecile Distances (Bimodal Distribution)

Decile	Point	Distance between Points
D_9	54.4	
		1.1
D_8	53.3	
		1.2
D_7	52.1	
		1.4
D_6	50.7	
		8.7
D_5	42.0	
		8.7
D_4	33.3	
		1.4
D_3	31.9	
		1.2
D_2	30.7	
		1.1
D_1	29.6	

Quartiles and Shape

- Symmetrical:
  - $Q_3 - Q_2 = Q_2 - Q_1$
- Positive Skew:
  - $Q_3 - Q_2 > Q_2 - Q_1$

- Negative Skew:
  - $Q_3 - Q_1 < Q_2 - Q_1$

Ogive Shape

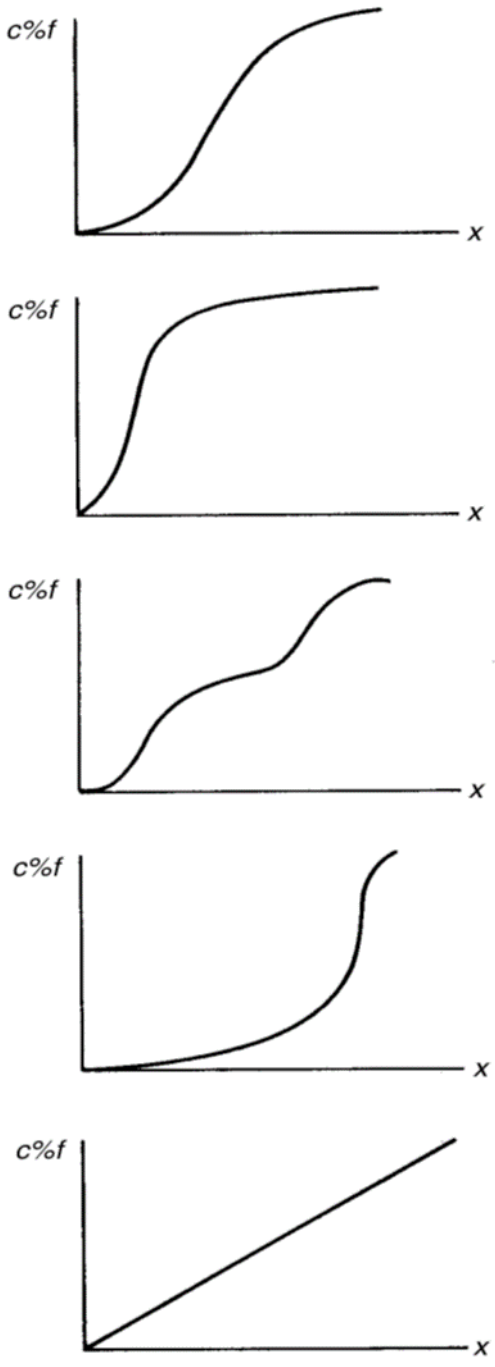


Figure 9: Ogive Shape

Calculating Percentiles and Percentile Ranks

- Percentiles:
  - $P_{\%} = X_{ll} + i\left(\frac{n(\frac{P_R}{100}) - \sum f_b}{f_i}\right)$
- Percentile Ranks
  - $P_R(X) = \frac{100}{n}\left(\sum f_b + \frac{f_i(P_{\%} - X_{ll})}{i}\right)$
    where:
  - X = Score of Interest
  - n = Total Sample size
  - $\sum f_b$ = Number of scores below interval containing X
  - i = interval size = $X_{ul} - X_{ll}$
  - X_{ll} = Lower limit of interval containing X
  - X_{ul} = Upper limit of interval containing X

Calculating Examples

- Using the table below, find $P_R(112)$, $P_R(158)$, P_{25}, P_{50}, P_{75}

```
##           freq cf
##              0  0
## [70,80)      0  0
## [80,90)      1  1
## [90,100)     6  7
## [100,110)    6 13
## [110,120)    2 15
## [120,130)    6 21
## [130,140)    4 25
## [140,150)    3 28
## [150,160)    0 28
## [160,170)    1 29
## [170,180)    1 30
```

Bar Graphs

- Bar graphs are a way to visually show a frequency table for qualitative variables.
- Can also be used to show other variables on the y-axis for qualitative variables.
- Has gaps to show the differences in groups.
- Order should be meaningful, either alphabetical for nominal variables or in the correct order for ordinal variables.

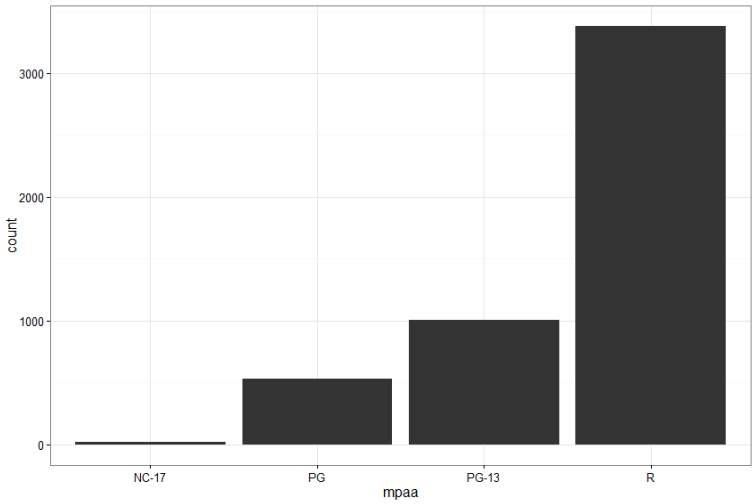


Figure 10: plot of chunk bar

Pie Charts

- Pie charts are useful to show the percentage of the whole for each group.
- The sum of the pieces of the pie must add up to 100% for this chart to be meaningful.

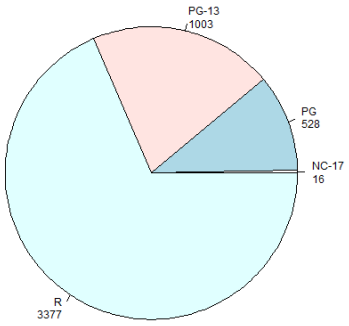


Figure 11: plot of chunk pie

Poor graphics

- Care needs to be made when constructing good graphics.

- It is easy for graphs to mislead/mask the purpose.
- The goal should be to easily convey the message.

Poor Charts

Interesting Graphic Examples

Languages other than English Spoken at home

The Racial Dot Map