Probability & Statistics Notes – Prof. Richard B. Goldstein

SOURCES OF DATA

Data may be collected in the laboratory, from economic measures, the Internet, or from files on a disk. The values may be given as individual values or already grouped into intervals.

GROUPING DATA INTO INTERVALS

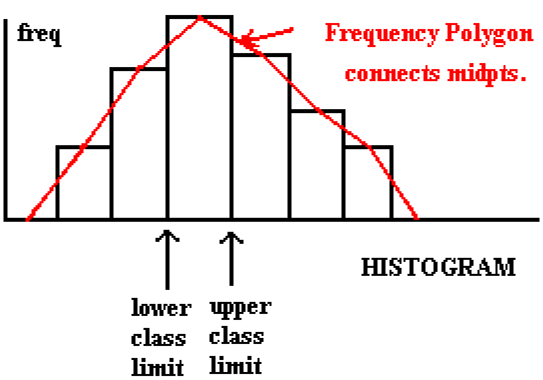
Simple rule: use 5 to 15 intervals depending upon the number of values and their numerical values

Strickberger: under 30 values – use 6 to 10
50 to 100 values – use 12
200 to 500 values – use 14

Martin: minimize the ratio: # sign reversals/ # of intervals

Although the interval sizes do not have to be equal, they are usually at worst simple multiples - for example, one or more intervals may be twice as wide as the others (if so, their bar heights should be halved).

HISTOGRAMS, FREQUENCY POLYGONS & OGIVES



data: $L = x_1 \le x_2 \le x_3 \le \cdots \le x_{n-1} \le x_n = H$

$$\boxed{\text{class width} = \frac{H - L}{\#\text{of intervals}}}$$

and is usually rounded up to the next integer

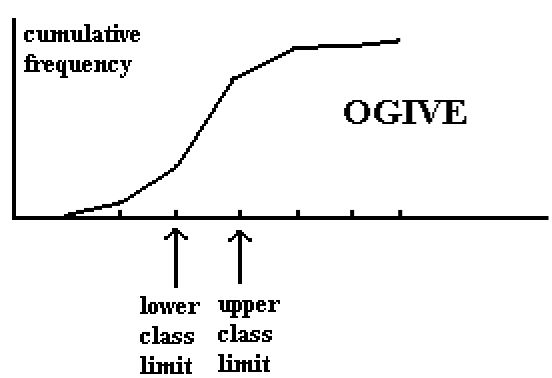
The frequency polygon connects the midpoints of each bar including one at zero on the left and right.

The bars must touch.

Each value fits into only one interval: $\boxed{\text{lower class limit} < \text{value} \le \text{upper class limit}}$

The cumulative frequency curve or ogive (pronounced "oh-jive") uses the same values on the x-axis as the histogram.

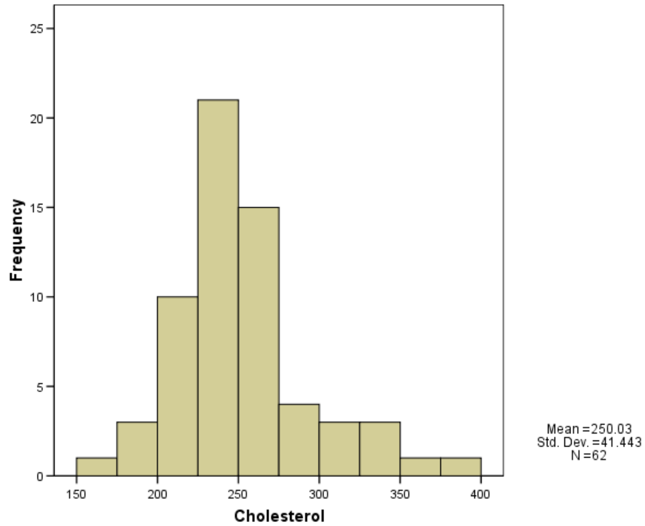
The shape is a non-decreasing curve or line segments from left to right and may use either the cumulative frequency on the y-axis scale from 0 to n or the cumulative percentage from 0% to 100%.



Cholesterol Data from the Framingham Heart Study

Examples: stem & leaf plot, histogram, Normal Q-Q plot, Box & Whisker Diagram with outliers (SPSS)

Stem-and-leaf plot		Freq	Cumul Freq
16	7	1	1
17		0	1
18	4	1	2
19	28	2	4
20	02	2	6
21	0125678	7	13
22	0556	4	17
23	0000122244668	13	30
24	03678	5	35
25	444668	6	41
26	347778	6	47
27	00288	5	52
28	35	2	54
29		0	54
30	008	3	57
31		0	57
32	7	1	58
33	46	2	60
34		0	60
35	3	1	61
36		0	61
37		0	61
38		0	61
39	3	1	62



Descriptives

			Statistic	Std. Error
Cholesterol	Mean		250.03	5.263
	95% Confidence Interval for Mean	Lower Bound	239.51	
		Upper Bound	260.56	
	5% Trimmed Mean		247.74	
	Median		241.50	
	Variance		1717.540	
	Std. Deviation		41.443	
	Minimum		167	
	Maximum		393	
	Range		226	
	Interquartile Range		44	
	Skewness		1.049	.304
	Kurtosis		1.816	.599

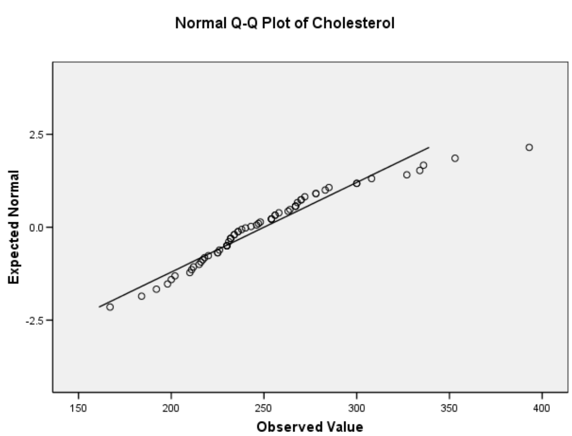
Percentiles

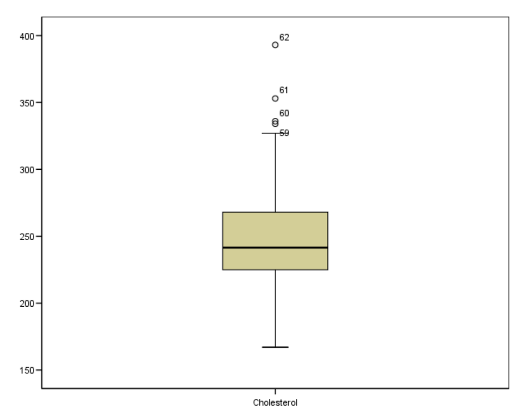
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Cholesterol	192.90	204.40	225.00	241.50	268.50	305.60	335.70
Tukey's Hinges	Cholesterol			225.00	241.50	268.00		

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Cholesterol	.105	62	.085	.939	62	.004

a. Lilliefors Significance Correction





<u>Moments and Percentiles</u> – Prof. Richard B. Goldstein

Discrete Sample Data: $x_1, x_2, \ldots, x_n$ **Ordered:** $L = x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} = H$

MEASURES OF CENTRAL TENDENCY

$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ is the sample **arithmetic <u>mean</u>**

$\tilde{x} = p_{50} = \begin{cases} \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \end{cases}$ is the sample **<u>median</u>** $\{L = x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} = H\}$

Trimmed mean cuts out a percentage of the data from each end

Weighted mean is $\dfrac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$

Geometric mean is $(x_1 x_2 \cdots x_n)^{1/n}$ if all $x_i > 0$

** **Harmonic mean** is $\dfrac{n}{\dfrac{1}{x_1} + \dfrac{1}{x_2} + \cdots + \dfrac{1}{x_n}}$

** μ_r is given by $\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n}$ **is the rth central moment about the mean**

MEASURES OF SPREAD

Variance and Standard Deviation

$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$ note: s^2 is an unbiased estimate

$s = \sqrt{s^2}$ is a biased estimate of the standard deviation

R = H – L is the **range**

** M.A.D. $= \dfrac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$ is the **mean absolute deviation**

IQR = Interquartile Range = $Q_3 - Q_1$

Chebychev's Theorem: For any $k \ge 1$ the proportion of the data that must lie within $\pm k$ standard deviations is *at least* $1 - \dfrac{1}{k^2}$ (ex. at least 75% of data within ±2 st. devs.)

** not in most texts

SKEWNESS (third moment) is a measure of the asymmetry of a distribution

Other measures include Pearson's skewness coefficient defined as $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

$$\hat{\alpha}_3 = \frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \quad \text{where } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{(used by Excel)}$$

** $\gamma_1 = \dfrac{\mu_3}{\mu_2^{3/2}}$

** Another Pearson measure of skewness involving the mode: $\dfrac{(\text{mean} - \text{mode})}{\text{standard deviation}}$

** Bowley's skewness defined as $\dfrac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \dfrac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$ using quartiles

KURTOSIS (fourth moment) is a measure of the peakedness of a distribution

$$\hat{\alpha}_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \text{(used by Excel)}$$

** $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$ and $\gamma_2 = \dfrac{\mu_4}{\mu_2^2} - 3$ is more common because it measures the excess from the normal distribution where $\beta_2 = 3$

PERCENTILES $p_k = k^{th}$ percentile

Note that the 80th percentile can be defined as either the lowest score that is "greater than" 80% of the scores or it can be defined as the lowest score "greater than or equal to" 80% of the scores. This can make a difference in small data sets.

Note that the k^{th} decile $d_k = p_{10k}$ and k^{th} quartile $Q_k = p_{25k}$. Also note that median = $\tilde{x} = p_{50} = d_5 = Q_2$

Consider the sorted sample: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

Method I (used by Excel and Quattro Pro for example)
note: The median will be the same for both methods

p_k = given by $x_{(r)}$ where $r = 1 + \frac{k}{100}(n-1)$
for example if n = 7 and k = 20, then $p_{20} = x_{(1 + 0.2(7-1))} = x_{(2.2)} = x_{(2)} + 0.2(x_{(3)} - x_{(2)})$
$x_{(k)}$ is in the $100\left(\dfrac{k-1}{n-1}\right)$ percentile
for example if n = 9 then $x_{(7)}$ is the $100(6/8) = 75^{th}$ percentile

Method II (used by SPSS and known as Tukey's Hinges)

p_k = given by $x_{(r)}$ where r = k(n + 1) with the following rules:

(a) if k(n + 1) < 1 then use r = 1
(b) if k(n + 1) > n then use r = n
(c) if k(n + 1) then interpolate as in Method I

$x_{(k)}$ is in the $100\left(\frac{k}{n+1}\right)$ percentile

Example: 4, 5, 8, 11, 15, 18, 19, 30

Method I - the 30th percentile $p_{30} = x_{(1+0.3(8-1))} = x_{(3.1)} = x_{(3)} + 0.1(x_{(4)} - x_{(3)}) = 8 + 0.1(11 - 8) = 8.3$
Method II – $p_{30} = x_{(0.3(8+1))} = x_{(2.7)} = x_{(2)} + 0.7(x_{(3)} - x_{(2)}) = 5 + 0.7(8 - 5) = 7.1$

Method I – 8 is in the 100(3 – 1)/(8 – 1) = 200/7 = 28.6th percentile
Method II – 8 is in the 100(3)/9 = 300/9 = 33.3rd percentile

SAMPLE CALCULATIONS

Sample Data: 3, 7, 10, 15, 18, 22, 37

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+7+10+15+18+22+37}{7} = \frac{112}{7} = 16$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(3-16)^2 + \cdots + (37-16)^2}{7-1} = \frac{768}{6} = 128 \text{ or } s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{2560 - 7(16)^2}{7-1} = \frac{768}{6} = 128$$

$$\hat{\alpha}_3 = \frac{n}{(n-1)(n-2)s^3} \sum (x_i - \bar{x})^3 = \frac{7}{6(5)128\sqrt{128}} \left[(3-16)^3 + \cdots (37-16)^3\right] = \frac{7(6342)}{3840\sqrt{128}} = 1.02185...$$

$$\hat{\alpha}_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4} \sum (x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)} = \frac{7(8)}{6(5)(4)128^2}\left[(3-16)^4 + \cdots (37-16)^4\right] - \frac{3(6)^2}{5(4)}$$

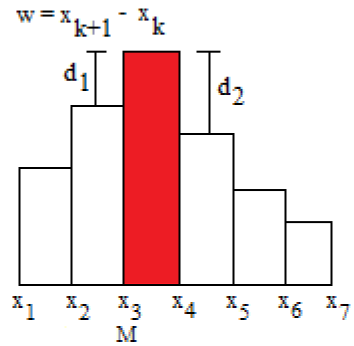
$$= \frac{56}{1{,}966{,}080}(232{,}212) - \frac{108}{20} = 6.614111... - 5.4 = 1.214111..$$

$p_{30} = x_{(1+0.3(7-1))} = x_{(2.8)} = x_{(2)} + 0.8(x_{(3)} - x_{(2)}) = 7 + 0.8(10 - 7) = 7 + 2.4 = 9.4$

Value	3	7	10	15	18	22	37
percentile	0	16.67	33.33	50.00	66.67	83.33	100

GROUPED DATA

Grouped Sample Data: x_1 with frequency f_1, x_2 with frequency f_2, …, x_k with frequency f_k



$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} \text{ where } n = \sum_{i=1}^{k} f_i$$ is the sample **arithmetic mean**

the **median** is found from the ogive

** the **mode** is either the largest class or more accurately $M + \left(\frac{d_1}{d_1 + d_2}\right)w$

where M is the left lower limit of the modal class and w is the common width

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n-1} = \frac{\sum_{i=1}^{k} x_i^2 f_i - n\bar{x}^2}{n-1}$$

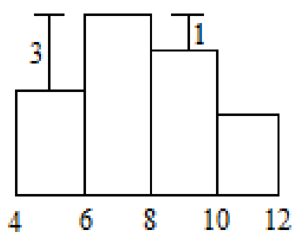
Percentiles are found by using the ogive (cumulative frequency curve) and interpolating.

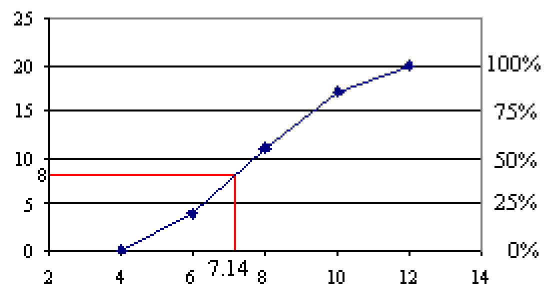
Consider the grouped sample data case

Example: class intervals 4 to 6, 6 to 8, 8 to 10, and 10 to 12

x_i	5	7	9	11
f_i	4	7	6	3

total frequency = $f_1 + f_2 + f_3 + f_2 = 20$





$$\text{Mean} = \frac{(5)(4) + (7)(7) + (9)(6) + (11)(3)}{4 + 7 + 6 + 3} = \frac{156}{20} = 7.8$$

$$\text{Mode} = 6 + \left(\frac{3}{3+1}\right)2 = 7.5 \qquad \text{Median} = 6 + \left(\frac{6}{7}\right)2 = 7.714$$

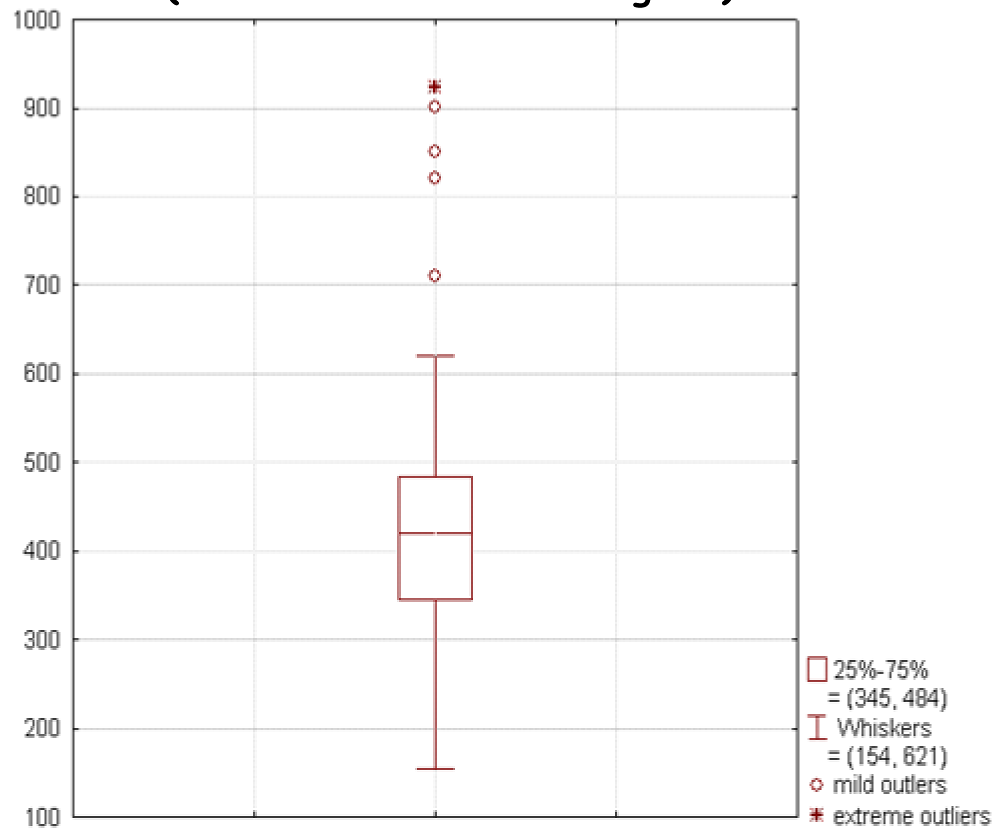
$$\text{Variance} = \frac{5^2 4 + 7^2 7 + 9^2 6 + 11^2 3 - 20(7.8)^2}{20 - 1} = \frac{75.2}{19} = 3.958 \quad \text{St Dev} = 1.989$$

$$\text{Pearson's Skewness} = \frac{3(7.8 - 7.714)}{1.989} = 0.1297 \text{ (slightly positive)}$$

p_{40} is at 0.4(20) = 8 on the y-axis and 6 + (4/7)2 = 7.143 on the x-axis

Medians and other percentiles are calculated using **<u>interpolation</u>** on the ogive

Box Plot (aka box-and-whisker diagram)



1		154
2		162
3		177
4		180
5		230
6		273
7		324
8	Q1 =	**345**
9		356
10		378
11		405
12		410
13		412
14		416
15	Q2 =	**420**
16		430
17		442
18		450
19		465
20		471
21		479
22	Q3 =	**484**
23		590
24		621
25		711
26		821
27		848
28		900
29		920

For this data set:

- Smallest non-outlier observation = 154
- Lower quartile $Q_1 = 345$
- Median $Q_2 = 420$
- Upper quartile $Q_3 = 484$
- Interquartile range $IQR = Q_3 – Q_1 = 484 – 345 = 139$
- Largest non-outlier observation = 621
- Mild outliers (o) are between 1.5*IQR and 3*IQR above Q_3 : (692.5, 901] and below Q_1 : [-72, 136.5)
- Extreme outliers (*) are above Q_3 + 3*IQR = 901 or below Q_1 - 3*IQR = -72
- The data is skewed to the right (positively skewed)

Rule for Whiskers:

The **lower whisker** starts at Q_1 and extends downward to Q_1 – 1.5(IQR) or the minimum value, whichever is greater.
The **upper whisker** starts at Q_3 and extends upward to Q_3 + 1.5(IQR) or the maximum value, whichever is lower.

References:

Exploratory Data Analysis, John W. Tukey, Addison-Wesley, Reading, MA 1977
Kendall & Stuart, *The Advanced Theory of Statistics*
Abramowitz & Stegun, *Handbook of Mathematical Functions*
http://mathworld.wolfram.com/Skewness.html and http://mathworld.wolfram.com/Kurtosis.html
http://www.answers.com/topic/skewness and http://www.answers.com/topic/kurtosis
http://en.wikipedia.org/wiki/Box_plot