

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

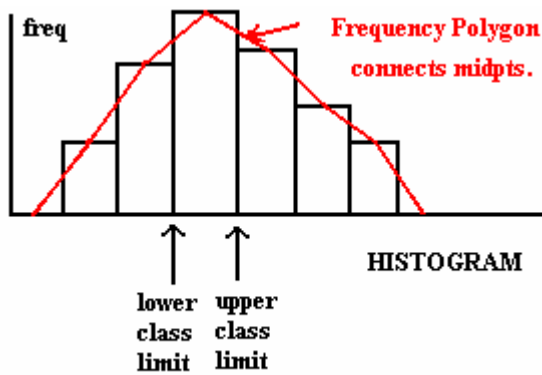
Simple rule: use 5 to 15 intervals depending upon the number of values and their numerical values

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: $L = x_1 \leq x_2 \leq x_3 \leq \dots \leq x_{n-1} \leq x_n = H$

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

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.

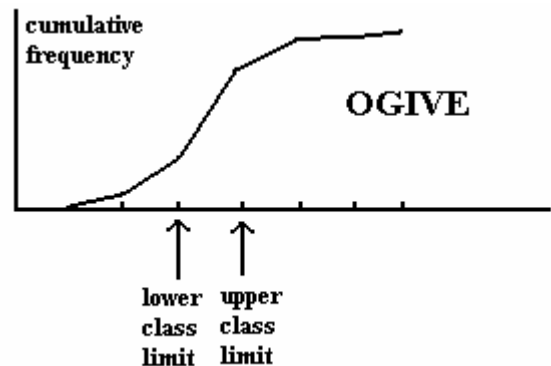
The bars must touch.

Each value fits into only one interval:

$$\text{lower class limit} < \text{value} \leq \text{upper class limit}$$

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

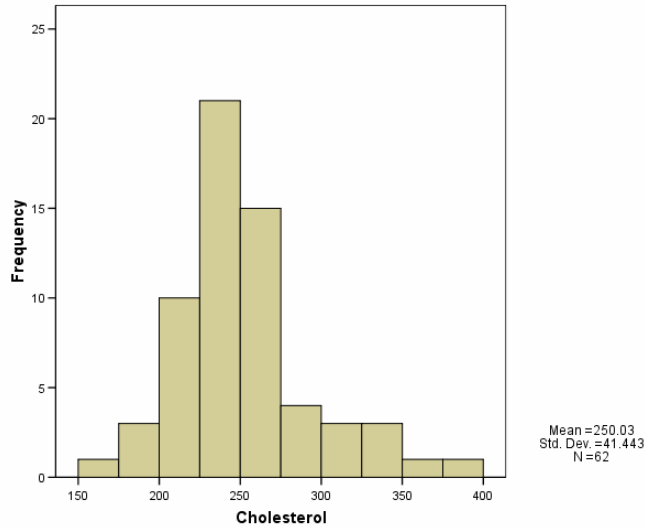
The shape is a non-decreasing curve or line segments from left to right and may use either the cumulative frequency on the y-axis scale from 0 to n or the cumulative percentage from 0% to 100%.



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

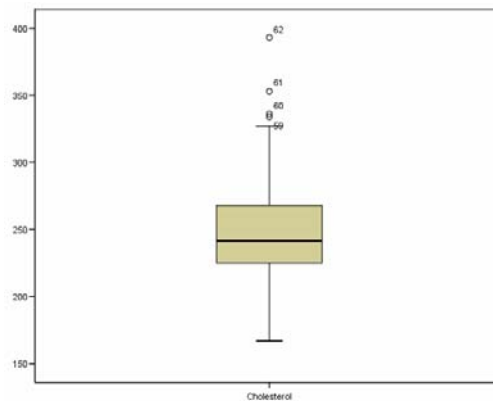
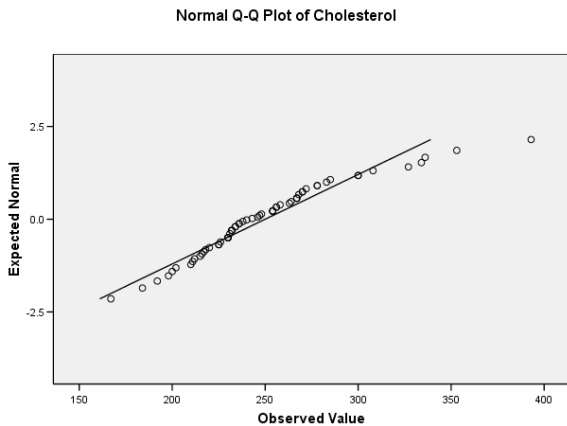


Descriptives			Statistic	Std. Error
Cholesterol	Mean		250.03	5.263
	95% Confidence Interval for Mean	Lower Bound	239.51	
		Upper Bound	260.56	
	5% Trimmed Mean		247.74	
	Median		241.50	
	Variance		1717.540	
	Std. Deviation		41.443	
	Minimum		167	
	Maximum		393	
	Range		226	
	Interquartile Range		44	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

		Percentiles						
		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		

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

a. Lilliefors Significance Correction



Moments and Percentiles

Discrete Sample Data: x_1, x_2, \dots, x_n

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

MEASURES OF CENTRAL TENDENCY

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

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

$\tilde{x} = p_{50} = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \end{cases}$ is the sample **median** $\{L = x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} = H\}$

Note: for grouped data the median is found from the ogive

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

Harmonic mean is $\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$

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

Weighted mean is $\frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}$

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

MEASURES OF SPREAD

Varaiance and Standard Deviation

$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ is an unbiased estimate of σ^2

$\frac{\sqrt{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s = \frac{s}{A_n}$ where $A_n \approx 1 - \frac{1}{4n-4}$ is an unbiased estimate of σ

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

is an unbiased estimate of σ^2 for grouped data

R = H – L is the **range**

$$\text{M.A.D.} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

is the **mean absolute deviation**

SKEWNESS (third moment) is a measure of the asymmetry of a distribution $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$

Other measures include Pearson's mode skewness defined as $\frac{(\text{mean} - \text{mode})}{\text{standard deviation}}$

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

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

An unbiased estimate for sample data is given by:

$$\hat{\alpha}_3 = \beta_1^2 = \text{est}(\gamma_1) = \frac{n}{(n-1)(n-2)s^3} \sum_{i=1}^n (x_i - \bar{x})^3 \text{ where } s^2 = \mu_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

or $\hat{\alpha}_3 = \text{est}(\gamma_1) = \frac{n}{(n-1)(n-2)s^3} \sum_{i=1}^k (x_i - \bar{x})^3 f_i$ where $s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 f_i}{n-1}$ for grouped data

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

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

An unbiased estimate for sample data is given by:

$$\hat{\alpha}_4 = \text{est}(\gamma_2) = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4} \sum_{i=1}^n (x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

or $\hat{\alpha}_4 = \text{est}(\gamma_2) = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4} \sum_{i=1}^k (x_i - \bar{x})^4 f_i - \frac{3(n-1)^2}{(n-2)(n-3)}$ for grouped data

References:

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>

PERCENTILES $p_k = k^{\text{th}}$ 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: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

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 $100\left(\frac{k-1}{n-1}\right)$ percentile

for example if $n = 9$ then $x_{(7)}$ is the $100(6/8) = 75^{\text{th}}$ 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 $p_{30} = x_{(1+0.3(8-1))} = x_{(3.1)} = x_{(3)} + 0.1(x_{(4)} - x_{(3)}) = 8 + 0.1(11 - 8) = 8.3$

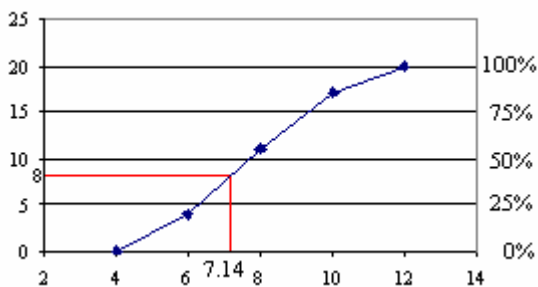
Method II - $p_{30} = x_{(0.3(8+1))} = x_{(2.7)} = x_{(2)} + 0.7(x_{(3)} - x_{(2)}) = 5 + 0.7(8 - 5) = 7.1$

Method I - 8 is in the $100(3-1)/(8-1) = 200/7 = 28.6^{\text{th}}$ percentile

Method II - 8 is in the $100(3)/9 = 300/9 = 33.3^{\text{rd}}$ percentile

Consider the grouped sample data case

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



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 = $f_1 + f_2 + f_3 + f_4 = 20$

p_{40} is at $0.4(20) = 8$ on the y-axis

and $6 + (4/7)2 = 7.143$ on the x-axis

SAMPLE CALCULATIONS

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

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

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

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

$$\begin{aligned} \hat{\alpha}_4 &= \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4} \sum (x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)} = \frac{7(8)}{6(5)(4)128^2} [(3-16)^4 + \dots + (37-16)^4] - \frac{3(6)^2}{5(4)} \\ &= \frac{56}{1,966,080} (232,212) - \frac{108}{20} = 6.614111\dots - 5.4 = 1.214111\dots \end{aligned}$$

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

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

Grouped Data

class mean x_i	class limits	frequency f_i	$x_i f_i$	$(x_i - m)^2 f_i$	$(x_i - m)^3 f_i$	$(x_i - m)^4 f_i$
3	2 to 4	2	6	18	-54	162
5	4 to 6	16	80	16	-16	16
7	6 to 8	8	56	8	8	8
9	8 to 10	3	27	27	81	243
11	10 to 12	1	11	25	125	625
sums		30	180	94	144	1054

$$\text{mean} = \bar{x} = m = \frac{180}{30} = 6$$

$$s^2 = \frac{94}{29} = 3.241379\dots \text{ and } s = 1.800383\dots$$

$$\hat{\alpha}_3 = \frac{30}{(29)(28)(1.800383\dots)^3} (144) = 0.91166\dots$$

$$\hat{\alpha}_4 = \frac{30(31)}{29(28)(27)(1.800383\dots)^4} (1054) - \frac{3(30-1)^2}{28(27)} = 4.255437\dots - 3.337302\dots = 0.918135\dots$$

percentile: p_{40} corresponds to the cumulative total of $0.40(30) = 12$
 by interpolation the x value is at $4 + 2 * (12 - 2) / 16 = 4 + 20/16 = 5.25$
 i.e, (lower class limit) + (class width) * (cumulative total - prior total) / class frequency