Chapter 8 - Sampling, Statistics, Data plots

Defn: A population consists of the totality of the observations with which we are concerned.

Defn: A sample is a subset of a population.

Each observation in a population is a value of a random variable with probability distribution $f(x)$

Defn: Let $X_1, X_2, \ldots X_n$ be n independent random variables each having the same prob. distribution $f(X)$. Define $X_1, \ldots X_n$ to be a random sample of size n from the population $f(x)$ and write its joint probability distribution as

$$\underbrace{f(x_1, x_2, \ldots x_n)}_{\text{Sample}} = f(x_1) f(x_2) \ldots f(x_n) = \prod_{k=1}^{n} f(x_k)$$

Defn: Any function of the random variables constituting a random sample is called a statistic.

Defn: Sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a statistic.

Important: Sample mean is not the same thing as the mean of a random variable but they are very closely related.

Defn: Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is a statistic.

Important: Again this is very related to the standard deviation of a random variable but is not the same thing.

Defn: Sample mode is the observation value that occurs the most number of times in a sample.

Example: (Exercise 8-4 textbook)

10 patients waited the following lengths of time (in minutes) in a doctors office:

5, 11, 9, 5, 10, 15, 6, 10, 5, 10

Sample mean: $\bar{X} = \frac{1}{10}(5+11+9+5+10+15+6+10+5+10)$

$= \cancel{8.6}\ 8.4$ minutes

Sample variance: $S^2 = \frac{1}{9}\Big((5-8.6)^2+(11-8.\cancel{6})^2+\dots$

$\dots+(10-8.6)^2+(5-8.\cancel{6})^2+(10-8.\cancel{6})^2\Big)$

$= \cancel{10.69}\ 10.933$

Sample standarddeviation is the square root of S^2.

so standard dev for example $\sqrt{10.93}$

Defn: Sample median is the middle value of a sample after sorting.

example: sort the observations

5, 5, 5, 6, 9, 10, 10, 10, 11, 15

$\uparrow\ \uparrow$

average = 9.5

If there is an even number of elements in the sample take the average of the $\frac{n}{2}$'th and $n/2+1$'th entry after sorting.

If there is an odd n, then simply use the $\frac{n+1}{2}$'th entry.

Mode: 5 occurs 3 times, 10 occurs 3 times, so the mode is 5 and 10.

When to use mean vs. median vs. mode to describe the center of the sample?

- Mean is sensitive to outliers. If there are large outliers we prefer the median over the mean.

- The mode is not desirable if the sample size is too small.

Example: (exercise 8.8 textbook) # sick days

Sample: 15, 7, 8, 95, 19, 12, 8, 22, 14

$\underline{\underline{95}}$ ↑ outlier

* Mean: $\overline{X} = 22.2$ (influenced strongly by the data point $x=95$)

* Median: Sort the data first.

7, 8, 8, 12, 14, 15, 19, 22, 95

$n=9$ median is the $\frac{9+1}{2} = 5$'th element $= 14$

$\tilde{X} = 14$.

Notice that the median is not influenced much by the outlier. The mean on the other hand is and comes out larger than 8 of the 9 observations in the sample. So we prefer the median in this case.

* The mode is 8 because 8 occurs the most times in the sample (specifically twice in this case). In a small sample like this ($n=9$) the mode is not very useful.

A faster way to compute the variance of a sample:

By definition $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i^2 - 2\bar{X} X_i + \bar{X}^2 \right)$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right]$$

substitute $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - \frac{2\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i\right)}{n} + \cancel{n} \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n^{\cancel{2}}} \right]$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2 \right]$$

$$\boxed{s^2 = \frac{1}{n(n-1)} \left[n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right]}$$

Box-and-Whisker Plot: A graphical tool to get an idea about the center, variability and degree of asymmetry of a sample.

Defn: A *quantile* of a sample, $q(f)$, is a value for which a specified fraction f of the data values is less than or equal to $q(f)$.

The sample median is $q(0.5)$

The 75th percentile is $q(0.75)$

The 25th percentile is $q(0.25)$

Example : (Example 8.3 textbook)

Sample :

1.09, 1.92, 2.31, 1.79, 2.28, 1.74, 1.47, 1.97,
0.85, 1.24, 1.58, 2.03, 1.70, 2.17, 2.55, 2.11,
1.86, 1.90, 1.68, 1.51, 1.64, 0.72, 1.69, 1.85,
1.82, 1.79, 2.46, 1.88, 2.08, 1.67, 1.37, 1.93,
1.40, 1.64, 2.09, 1.75, 1.63, 2.37, 1.75, 1.69

How to compute the median, 25th and 75th percentiles?

Sample size $n = 40$. Sort the sample (easiest way to do this is with MATLAB's sort command) : Xs = sort(X) where X is an array containing the sample.

Then the 25th percentile is the $\lceil \frac{25 \times n}{100} \rceil$'th element in the sorted list.

$\lceil x \rceil$ means round x up to the nearest integer

In our case $\frac{25 \times n}{100} = \frac{25 \times 40}{100} = 10$ so $q(.25) = Xs(10) = 1.63$

see below

Sorted array :

0.72	0.85	1.09	1.24	1.37	1.40	1.47	1.51
1.58	1.63	1.64	1.64	1.67	1.68	1.69	1.69
1.7	1.74	1.75	1.75	1.79	1.79	1.82	1.85
1.86	1.88	1.9	1.92	1.93	1.97	2.03	2.08
2.09	2.11	2.17	2.28	2.31	2.37	2.46	2.55

25% percentile
10th element

median

30th element
75% percentile

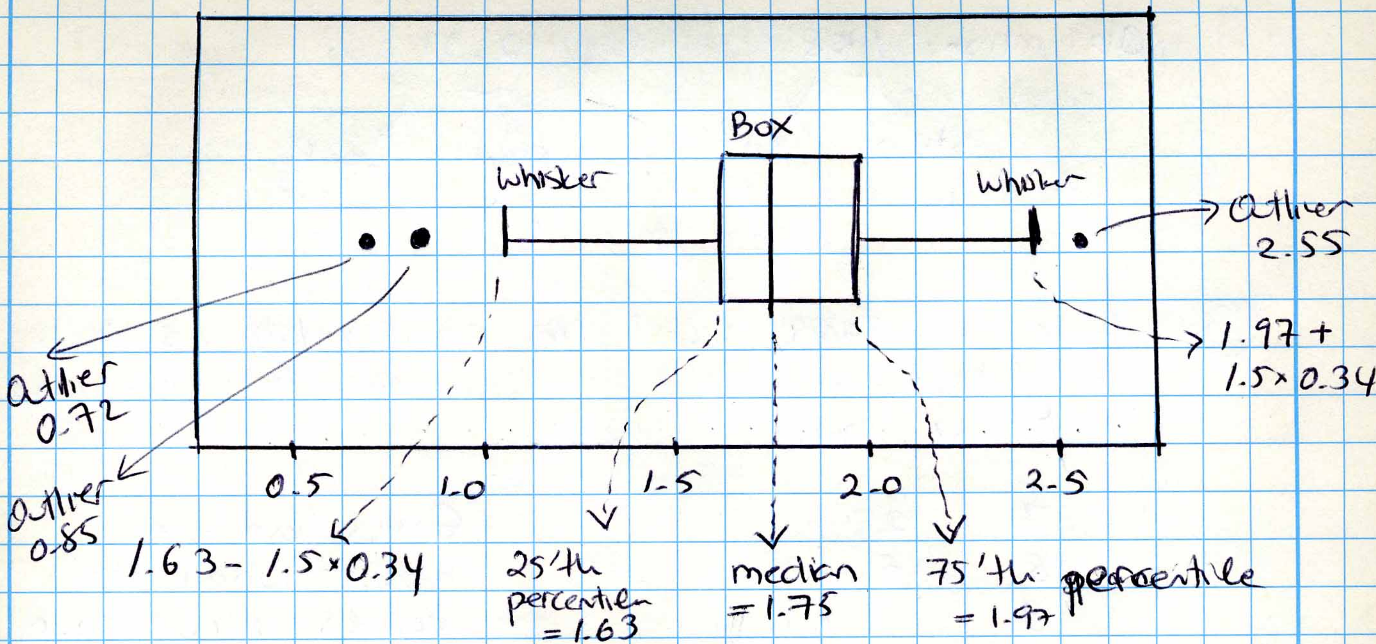
Similarly the median is the $\frac{50 \times 40}{100} = 20$'th element

$q(0.5) = Xs(20) = 1.75$

Finally, the 75th quantile is $\frac{75 \times 40}{100} = 30$'th element

$q(0.75) = Xs(30) = 1.97$

Box-and-Whisker Plot

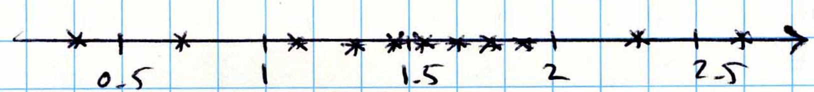


Defn: Interquartile range: $q(0.75) - q(0.25)$

In our example $1.97 - 1.63 = 0.34$

The whiskers are drawn at a distance of 1.5 times the interquartile range from the 25'th and 75'th percentiles. Anything outside that range is shown as an outlier.

Why not simply plot the individual observations along the x-axis?



Hard to interpret the data or to get an idea about the density function this way. Although in 2D (joint r.vs) the scatter plot is useful.

<u>Another graphical tool</u>: Stem-and-leaf plot

(1) Split each observation ~~[illegible]~~ INTO 2 parts: stem and leaf. Example: stem can be the digit preceeding the decimal and the leaf the digits after the decimal. In our example better to have the stem be the digits before and immediately after the decimal.

Observation: 1.58 , 2.31

Stem 15 Leaf 8

Stem 23 Leaf 1

(2) Make a table: List the stem values as rows. Add each leaf value with a specific stem value to that row.

Stems	Leafs
7	2
8	5
9	
10	9
11	
12	4
13	7
14	0 7
15	1 8
16	3 4 4 7 8 9 9
17	0 4 5 5 9 9
18	2 5 6 8
19	0 2 3 7
20	3 8 9
21	1 7
22	8
23	1 7
24	6
25	5

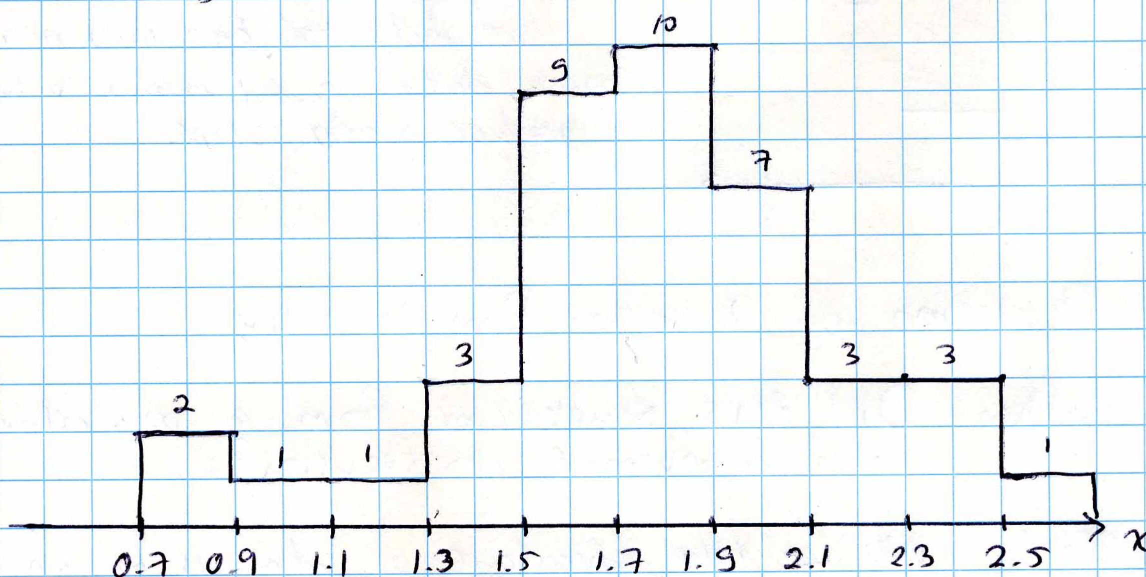
Gives an idea about what stem values occur more frequently. In this case, we could guess that the density function for the random variable peaks around 1.6 – 1.7 range.

Remember: All observations in a sample are drawn from a density function (population) which we may or may not know.

Another graphical tool: Histogram: A histogram is very much like a stem-and-leaf plot in terms of the information it provides. Define some bins (range of values), count how many observations fall in each bin, and make a bar graph.

Example: For our data lets choose the bins

		Count
Bin 1 :	$0.7 \le X < 0.9$	2
Bin 2 :	$0.9 \le X < 1.1$	1
Bin 3 :	$1.1 \le X < 1.3$	1
Bin 4 :	$1.3 \le X < 1.5$	3
Bin 5 :	$1.5 \le X < 1.7$	9
Bin 6 :	$1.7 \le X < 1.9$	10
Bin 7 :	$1.9 \le X < 2.1$	7
Bin 8 :	$2.1 \le X < 2.3$	3
Bin 9 :	$2.3 \le X < 2.5$	3
Bin 10 :	$2.5 \le X$	1



Normally bins are equally spaced and cover the entire range of observations.

Histograms in MATLAB: 2 commands hist, histc

hist (x , 20) covers the entire range of x with 20 bins

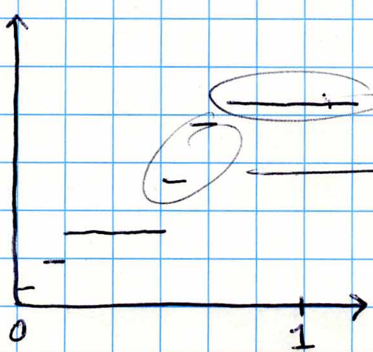
hist (x, a) where "a" is an array like [0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4]

uses the elements of "a" as bin centers

histc (x, a) treats the elements of "a" as edges of bins

* A quantile plot simply plots the data values on the vertical axis agaist its quantile on the horizontal axis.

Let y_i be the ith observation when they are sorted low to high. Then y_i is the i/n'th quantile where n is the size of the sample. So we plot y_i vs. i/n. For theoretical reasons we normally plot y_i vs. $f_i = \dfrac{i - 3/8}{n + 1/4}$



→ Large clusters around specific values are indicated by slopes near 0

→ Sparse data around certain values produce steep slopes.

* Detection of Deviation from Normality.

Question: Did this sample come from a poplation with a normal distribution?

Tool. We can take advantage of what is known about the quantiles of the normal distribution to answer this question.

Approximation to f'th quantile of $N(\mu, \sigma)$:

$$q_{\mu,\sigma}(f) \approx \mu + \sigma\left[4.91\left(f^{0.14} - (1-f)^{0.14}\right)\right]$$

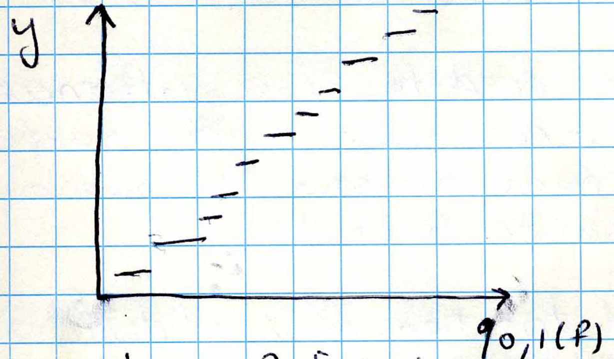
For the standard normal distribution, this simplifies to

$$q_{0,1}(f) \approx 4.91\left(f^{0.14} - (1-f)^{0.14}\right)$$

<u>Defn</u>: The normal-quantile-plot is a plot of the ordered observations y_i agaist $q_{0,1}(f_i)$ where $f_i = \dfrac{i - 3/8}{n + 1/4}$.

Data from Normal Dist

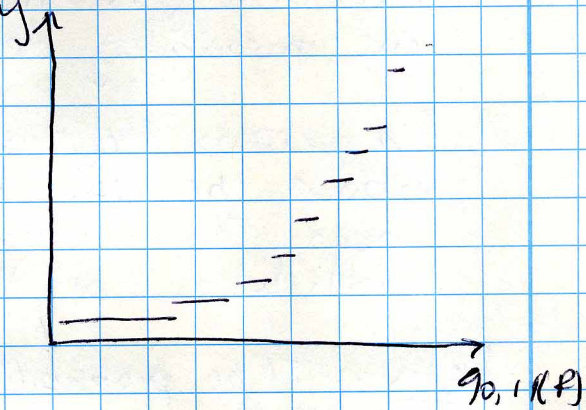
Straight-line relationship (approximately)

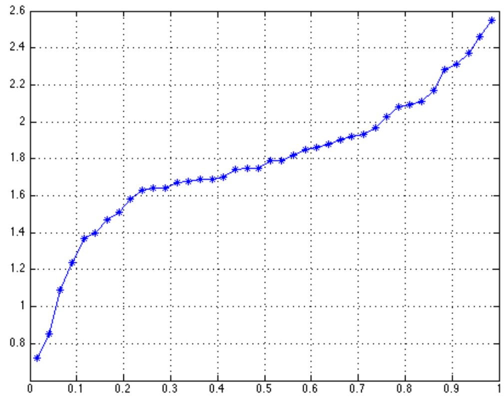


slope of line is an estimate of σ of the normal distribution. Intercept on the vertical axis is an estimate of μ.

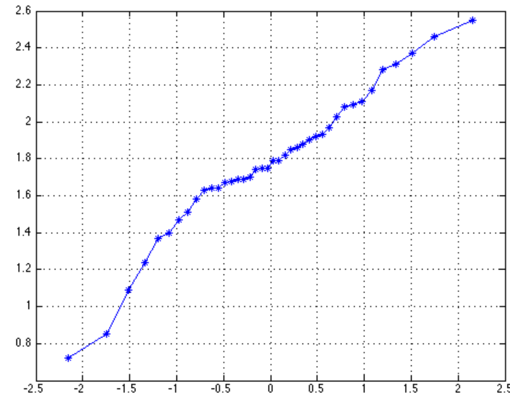
Data not from Normal Dist

Non-linear relationship

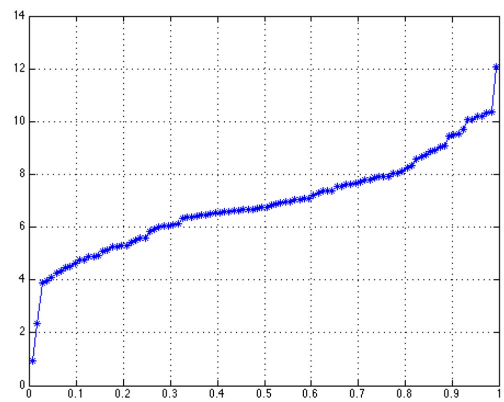




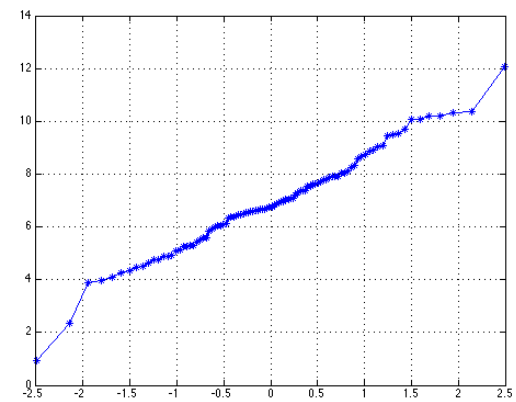
Example 8.3 Quantile plot



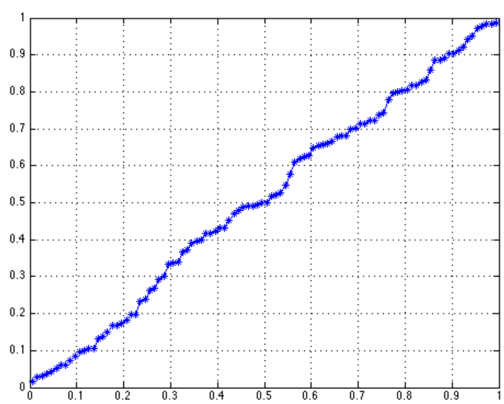
Example 8.3 Normal Quantile plot



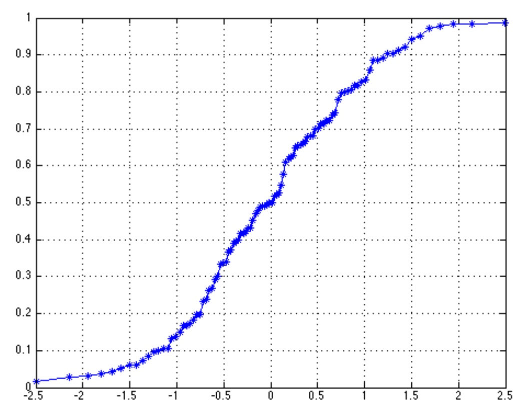
$N(\mu = 7, \sigma = 2)$ Quantile plot



$N(\mu = 7, \sigma = 2)$ Normal Quantile plot



Uniform dist 0 to 1 Quantile plot



Uniform dist Normal Quantile plot

The first row of figures correspond to the data given in Example 8.3 of the textbook. The second raw corresponds to a sample of 100 observations drawn from a normal distribution ($\mu = 7$ and $\sigma = 2$). The last row corresponds to a sample of 100 observations drawn from an uniform distribution.