.1 1/2 CH. 6. RANDOM SAMPLING AND DATA DESCRIPTION

6.1 DATA SUMMARY AND DISPLAY

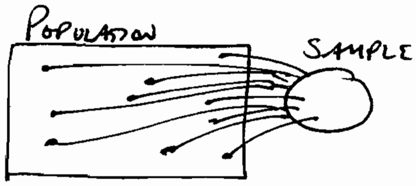
RECALL: LINEAR COMBINATIONS OF IID RV^S WITH μ AND σ^2

EX $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$.

$E(\bar{x}) = \mu$, $Var(\bar{x}) = \sigma^2/n$

MOREOVER, IF $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

THEN $\bar{X} \sim N(\mu, \sigma^2/n)$



SUPPOSE WE ARE INTERESTED IN THE MEAN OF A POPULATION.

SINCE WE CAN'T MEASURE EVERY ELEMENT,

TAKE SAMPLE OF SIZE n AND USE SAMPLE STATISTICS TO ESTIMATE POPULATION PARAMETERS.

BECAUSE $E(\bar{X}) = \mu$, $\bar{X}$ GIVES A REASONABLE ESTIMATE OF μ, THE POPULATION MEAN.

$\bar{X} \neq \mu$ IF WE KNOW THE DISTRIBUTION THEN WE CAN TELL

$E(\bar{x}) = \mu$ HOW "GOOD" THE ESTIMATE IS.

.2

	MEAN	STD. DEV
POPULATION PARAMETERS	μ	σ
SAMPLE STATISTIC	$\bar{X}$	S

2/2

Numerical Summaries

EX The number of earthquakes of magnitude 7 or greater for years 1980–99.

18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23

1) Measures of Center.

a) Mean $= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\bar{X} = \frac{18+14+\cdots+23}{20} = 14.75$$

b) Median = ~~~~ ½ way point of ordered data

6, 7, 8, 8, 10, 11, 11, 12, 14, 15, 15, 15, 16, 16, 18, 20, 22, 23, 23, 25

$$\frac{15+15}{2} = 15 = M$$

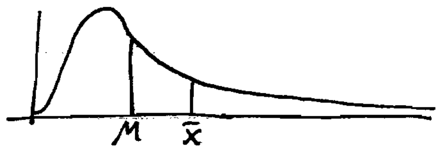
c) Mode = Most common data point. Mode = 15 since 3 15's.

If data highly skewed, usually use median as center

Note: a) Location of median is $\frac{n+1}{2}$ of sorted data.

b) If data is skewed, mean is pulled to the tail.

Can get an idea of how skewed by comparing $\bar{X}$ and M



2) Measures of Spread (Variability)

a) Sample Variance $= S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

"Degrees of Freedom" df,

so $E(s^2) = \sigma^2$ — Lose one for estimating ~~x~~ μ with $\bar{X}$.

Sample StDev $= s = \sqrt{s^2} = 5.72$.

b) Five-number Summary

min	Q_1	M	Q_3	max
6	10.5	15	19	25
	median of 1st ½ of data		median of 2nd ½ of data	

c) Range = max − min = 25 − 6 = 19

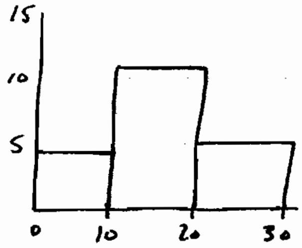
d) Interquartile Range = IQR = $Q_3 - Q_1 = 19 - 10.5 = 8.5$

4.3.5 3/3

Graphical Summaries

6.4 1) **Histograms** – Create categories (or bins) and count frequency of observations in each

Categories	Count
0–9	4
10–19	11
20–29	5



Categories	Count
0–4	0
5–9	4
10–14	5
15–19	6
20–24	4
25–29	1

Both: symmetric, unimodal

6.3 2) **Stem-and-Leaf Plot** (small dataset histogram)

Leaf is the last digit, all else is stem (may want to round first).

Stems	Leaves
0	6 7 8 8
1	0 1 1 2 4 5 5 5 6 6 8
2	0 2 3 3 5

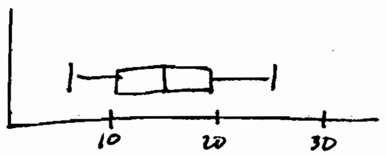
Split-stem — Most split into equal parts.

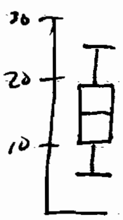
	Stems	Leaves
(0–4)	0	
(5–9)	0	6 7 8 8
(10–14)	1	0 1 1 2 4
(15–19)	1	5 5 5 6 6 8
(20–24)	2	0 2 3 3
(25–29)	2	5

Good for small data sets.

6.5 3) **Box-Plot** – Plot of five-number summary

Min	Q1	M	Q3	Max
6	10.5	15	19	25





Often programs plot whiskers out to $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$, after which they designate outliers with * * stars.

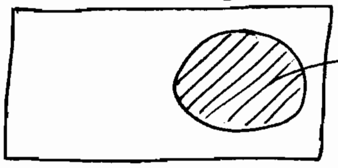
$$Q_1 - 1.5(IQR) = 10.5 - 1.5(8.5) = 10.5 - 12.75 = -2.25$$

$$Q_3 + 1.5(IQR) = 19 + 12.75 = 31.75$$ Beyond these points would be outliers

7.4 1/1

7.4 Sampling Distributions

Population of interest
with unknown μ

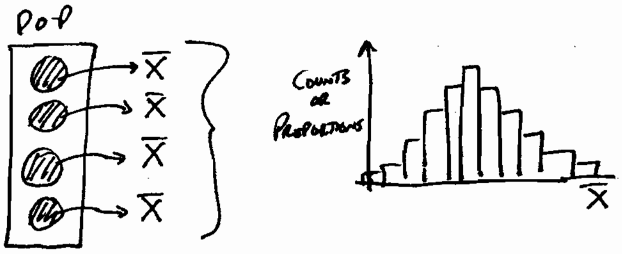


Sample which gives
information about population.

For example $\bar{X}$ estimates μ.

Notes

1) $\bar{X}$ estimates μ, but there is error associated with it.
2) Different samples give different estimates of μ.
3) Before collecting the ~~sample~~ data, we can think of $\bar{X}$ as a RV.
   Imagine listing all possible samples of size n and the $\bar{X}$s that result, plotting them with a histogram.



Histogram of possible $\bar{X}$s.

This gives the Sampling Distribution of $\bar{X}$,
which is basically the probability distribution of $\bar{X}$.

~~Denote~~: $\cancel{E(\bar{x}) = \mu_{\bar{x}} = \mu}$

$\cancel{Var(\bar{x}) = \sigma^2_{\bar{x}} = \sigma^2/n}$

~~If the population has a normal distribution, so does $\bar{X}$.~~

7.5 1/1

7.5 Sampling Distribution of Means, $\bar{X}$

Denote: $E(\bar{x}) = \mu_{\bar{x}} = \mu$

$Var(\bar{x}) = \sigma^2_{\bar{x}} = \sigma^2/n$

If the Population has a Normal Distribution, so does $\bar{X}$.

What type of Distribution does $\bar{X}$ have if the Population Distribution is NOT Normal?

[Handout 6] Dice: Population Distribution, $X \sim$ Disc Unif(1/6)

By the Central Limit Theorem (CLT), for n large enough,

$\bar{X} \overset{a}{\sim} N(\mu, \sigma^2/n)$ approximately regardless of the original Distribution.

Similarly, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$, as $n \to \infty$.

Handout 6b (Results of Handout 6) $X_1, .., X_{25} \overset{iid}{\sim}$ Disc Unif (1/6)

$E(X) = \mu = 3.5$ $E(\bar{x}) = \mu_{\bar{x}} = \mu = 3.5$

$Std(X) = \sigma = 1.708$ $Std(\bar{x}) = \sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.3416$

Observed: Mean $= \bar{x} =$

StDev $= s_{\bar{x}} = s/\sqrt{n} =$

Handout 6b: $\bar{X}$ has Normal Distribution.