1/20/15 263 Lecture 2

- Quiz 1 on Thurs
- Website is updated w/ Excel HW 1
- Excel HW 1 is DUE Thurs in class or in office (M702) Before 5:30.
- Webassign for Ch1 DUE Fri.

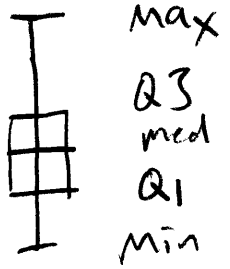
Last Time:

1.1 - 1.3 Data Sets, Histograms, Summary Statistics.

Data set: Cases → a particular instance
Variables → Quantities of ea. case eg "name" "height"

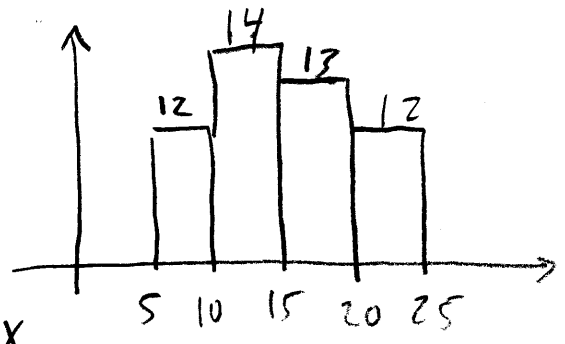
Distribution of a Variable:
The values taken & how often they are taken.

5-#-Summary: of a distribution



1.5 IQR rule for outliers

Histogram:
Choose bins w/ equal width, & count # of data points with Variable X in each bin.



Mean & Std Deviation:

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

"ADD values & Divide by number of data pts"

Variance $$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

"Averaged (squared) Distance from data pts to Mean"

Std Deviation $$s = \sqrt{s^2}$$

* why squared? generalization of Eucidean (Pythagorean) Distance, want to eliminate ± sign.
* why divide by n-1 instead of N? Knowing the mean reduces "degrees of freedom" by 1. It's more accurate! (See Wikipedia "Bessel's correction")
* why s instead of s^2? s will have the same <u>units</u> as the variable X.

Robustness of $\bar{X}$ & s:

Because each data pt is given "equal weight", mean & Var/std dev are <u>sensitive</u> to outliers. Q1, Q2, etc are not!

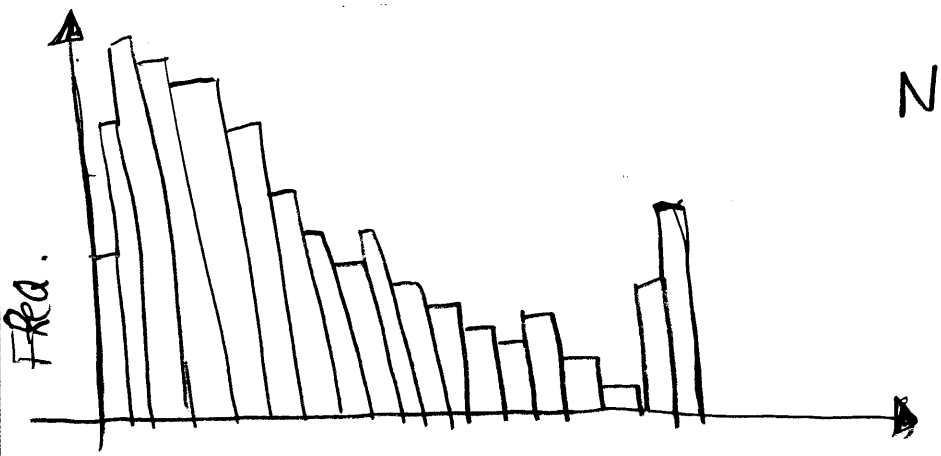
1.4: Normal Distribution, z-values

Note: Quantitative Variables can either take a <u>discrete</u> # of values, e.g. Class size, SAT Score, # of bacteria -Or- a <u>Continuous</u> # of values i.e. exact length, Sea level, temperature.

Occasionally, we will use <u>Continuous</u> vars as a <u>model</u> for discrete ones. Why? <u>Calculus</u>. (Note: Calc is not req'd for 265!)

Again: a distribution shows the values that a variable takes on the Horizontal axis, & the frequency taken by those values on the Vertical axis.

<u>Ex</u> Income Distribution, from US Census.



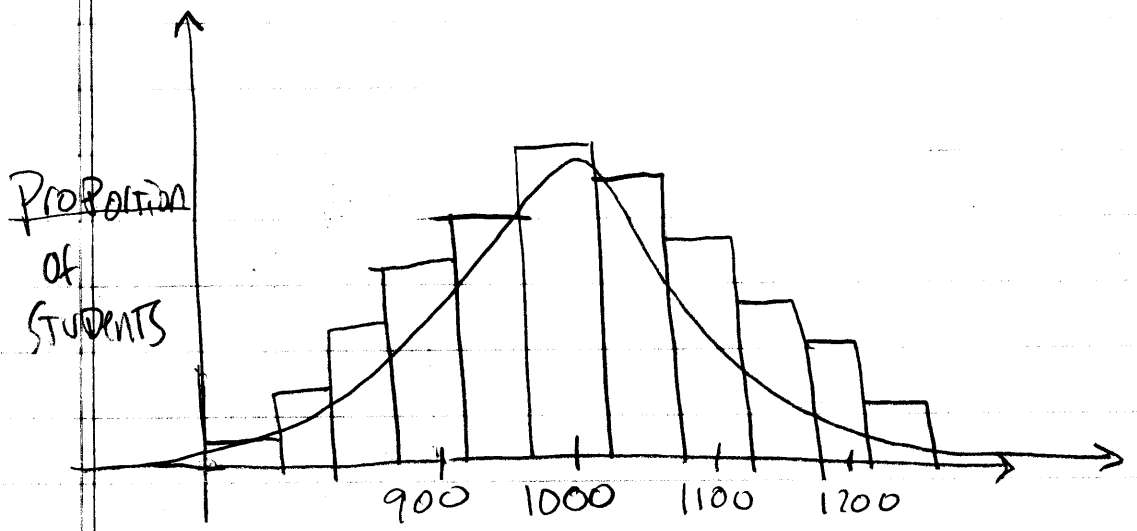
Note the Shape - Skewed "Right" (long Right tail)

Units:

*Horizontal axis is Income

*Vertical axis is # of people -or- Proportion of Population.

Density curves:



Mean: $\bar{x} \approx 1000$

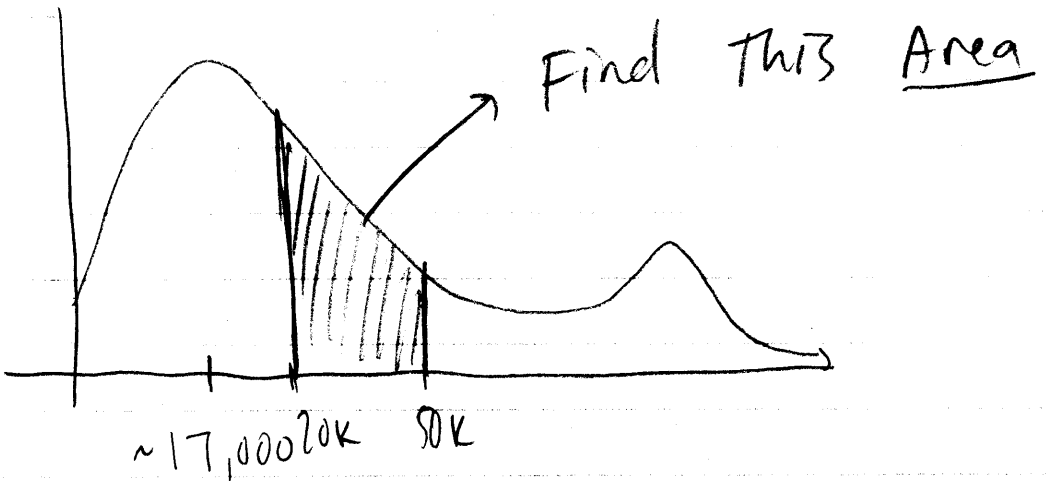
Std Dev: $s \approx 100$

Notice Density curve is Symmetric, Unimodal (one Peak), has a mean of 1000, Std Dev ~ 100. "Bell-Shaped"

- getting a high score seems as common as getting a low score
- many scores are bunched around a single score
- a large Portion of scores Are Concentrated w/in 2-3 std devs from $\bar{x}$

How will we use distributions?

- Want to answer questions like:
"What proportion of the population earns an income between $20k & $50k?"



Note: For a discrete variable, we can ask questions like "How many classrooms have 32 students?"

For a continuous variable, we can only ask for an interval of values "How many days have a temp between 57°F & 72°F?

For Calculus-minded people: We are computing integrals of distributions.

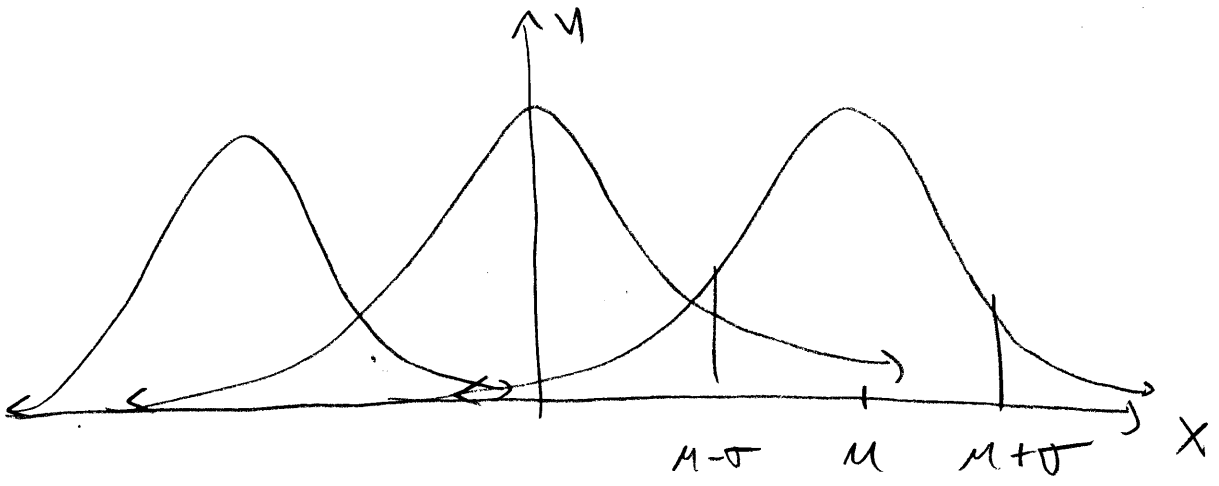
Normal distributions

A "normal" or "Gaussian" distribution is the bell-shaped curve w/ graph

"$N(\mu, \sigma)$" $\quad y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Two parameters describe it: μ ~ mean
σ ~ STD Dev.

Because of the <u>central limit theorem</u>, normal distributions are very common. We will return to this in chapter 5.



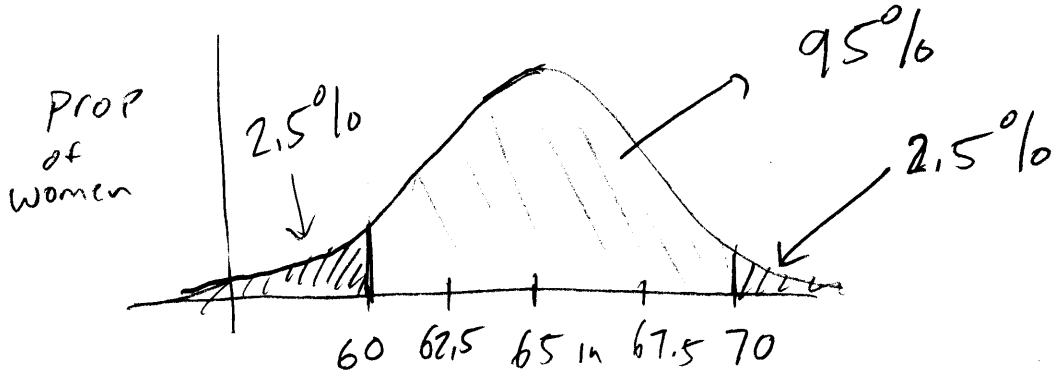
Properties

- total area is 1 ("100%")
- 68-95-99.7 rule:
  68% of data is within 1 std dev of mean
  90% " " " " 2 " "
  99.7% " " 3 " "

Normal Distribution examples

Suppose Women's heights are normally distrib. w/ mean 65 inches & std dev 2.5 in.

a) What prop. of women are less than 60 in?



b) What prop are less than 70 in?
$95\% + 2.5\% = 97.5\%$

c) What prop. are more than 72 in?

Z-Scores

We "change units" so we can always use $N(0,1)$:

$$z = \frac{x - \mu}{\sigma}$$

units of z are "standard deviations away from mean".
IE If $z = 1.5$, x is 1.5 σ's to the right of μ. If $z = -1.6$, x is 1.6 σ's to the left of μ.