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 (M 102) before 5:30.
- Webassign for Ch 1 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 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)
\[ s^2 = \frac{1}{N-1} \sum_{i=1}^{2} (x_i - \bar{x})^2 \] (Averaged squared distance from data pts to mean)
\[ s = \sqrt{s^2} \]

* Why squared? Generalization of Euclidean (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 units as the variable X.

Robustness of \( \bar{x} \) & S:
Because each data pt is given "equal weight," mean & var/3rd dev are sensitive to outliers, Q1, Q2, etc are not.
1.4: Normal Distribution, Z-Values

Note: Quantitative Variables can either take a **discrete** # of values, e.g., Class size, SAT Score, # of bacteria - or a **continuous** # of values i.e. exact length, Sea level, temperature.

Occasionally, we will use **continuous vars** as a model for **discrete ones**. Why? **Calculus**. (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.

**Ex 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:

![Histogram and curve]

Total SAT score (1600 max)

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 \( \approx 100 \). "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 within 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?"

Find this area:

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 with graph

\[ N(\mu, \sigma) \quad y = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \]

Two parameters describe it: \( \mu \) mean and \( \sigma \) std dev.

Because of the Central Limit Theorem, 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% is within 2 std dev of mean
  - 99.7% is within 3 std dev of mean
Normal Distribution Examples

Suppose women's heights are normally distributed with mean 65 inches and 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 \( \sigma \)'s to the right of \( \mu \). If \( Z = -1.6 \), \( X \) is 1.6 \( \sigma \)'s to the left of \( \mu \).