Statistical Summaries of Data

Review

Manufacturing process

– Process variation
  • Mixing the chips in the dough
    Hard to control, though smaller chips help.
  • Packing the cookies into bags (21-26/pouch)

Measurement process

– Definitions, measurement error
  • What is a chip?
  • How to count them once the cookie is made?
  • Role for *replication* in study design:
    If we only observed one cookie for each team, would the differences among teams be due to the definitions, or due to the natural variation in the process?

Variation and cost

– Low variation is expensive to run in the cookie example, but

– Saves money in lowering material costs.

Administrative Details

TAs in SH-DH 3009, *Web page*

Key Applications

Ten-minute summary of a large set of data

– Graphical vs. numerical

Daily earnings at risk in financial investments

– Normality, probabilities and risk aversion

– Will see more of this in Class 3

Definitions (Terminology)

Location (the center of the data)

– mean, average, first moment of inertia

– median, 50th percentile (quantiles)

– "trimmed" mean (athletic judging)

Scale (the dispersion of the data)

– variance and standard deviation (SD)
Note that a variance is itself an average, an average of *squared* distances rather than the original data.

– interquartile range (IQR)

Shape

– What's left over after remove numbers from plot axes

– E.g.:
Skewness versus symmetry, one mode versus two.

Outliers

– Unusual or aberrant values

– Impact on location and scale

Concepts

Normal distribution

– Bell-shaped curve

– Identified by two unknown values (called parameters)
mean `m` and SD `s`

Empirical Rule

– Normality + (mean, SD) –> probabilities

– $^{2}/_{3}$ within ± 1 SD

– $^{19}/_{20}$ within ± 2 SD's

Quantile plot

– Diagnostic for checking the validity of normality assumption

– Can't recognize normality from a histogram.
You can explore this claim using JMP-IN to simulate data from a normal population. Would you recognize the histograms from these samples as normal, or do they seem to lack the bell-shaped form?

Discussion

Population parameter versus sample statistic

Greek symbols and common notation.

Sample estimates of parameters

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}, \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Variance versus SD

Examples for Today

GMAT scores of the Wharton class of 1994 (page 9)

– What are typical GMAT scores for Wharton?

– Plots: histogram and boxplot

– Smoothing: kernel density, number of modes – use the "slider"

– Diagnostic for empirical rule: normal quantile plot

– Empirical rule works in the middle, but not extremes

Returns on General Motors stock (page 23)

– How has GM's stock done?

– Time series: sequence plots, trends/dependence

– Passage of time as a source of variation

– Use of histograms for time series

– Transformation 1: relative changes versus prices
Original data is not normal, transformed data is.

– Outliers (GM87 data set)

Skewness in executive compensation (pages 34, 45)

– What are typical incomes for top executives?

– Outliers or skewness?

– Transformation 2: logs for skewed, non-normal data

– Interpretation of log transformation

– Grouping (by industry) as a source of variation

– How did Eisner do in 1998… $589 Million.