Standard Error

Review

Variance

– One of many possible measures of the *variability* in data.
An alternative is the length of the box in the boxplot (IQR).

– How is variance defined?
In the language/notation of random variables…

$$Var(Y) = \sigma^2 = E(Y-\mu)^2, \quad \mu = E\,Y$$

– How is variance estimated from data?
As a sample estimate from data…

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

Sources of variation (See Class 3 of *Basic Business Statistics*)

– Grouping as a source of variation

– Time as a source of variation

– Value of transformations (e.g. logs) in revealing/exposing variation

– Process variation vs. measurement variation (cookies)

Probability models (Class 3 of course pack)

– Random variable relates outcomes to probabilities.

– Role of a model: Extrapolation to other situations
Compute the probability of an event that does not appear in the data.

– Sensitivity to assumptions

– Matching/calibrating model to observations

– t model versus normal model
Neither model is likely the "right model" for the data, but both provide reasonable descriptions of the observed variation. They differ, however, in the inferred probability of an extreme event. The normal model suggests a 10% drop is uncommon (p = 0.0021) whereas a t-model makes it more expected (p=0.0224).

Quantile plots

– Most important is knowing when to use/interpret these.

– Bands convey the ambiguity of the data:
Many models can be used to describe the variation in the data. The Bands are quite wide when one only observes a small sample, and remain quite wide in the extremes for large samples as well.

Administrative Details

Name tents

– Please bring to every class

Feedback forms

– Please fill in one for "your" class or any other.

Assignment, Cases

– Need to do statistics in order to learn statistics.

– JMP-IN review for Assignment #1 is tonight.

JMP-IN questions?

Column descriptors

Note that a column may be described as "continuous", "ordinal", or "nominal". The choice of these is important; JMP produces different analyses depending on how the data are characterized by these labels.

Key Applications for Today

Quality control

– Capable process versus an in-control process

– Setting guidelines for monitoring a process

– Tradeoffs: catching quickly versus "false positives"

– Tracking averages instead of individual observations

Risk of financial investments

– Use same tools to monitor investments rather than manufacturing

Definitions

Capable process

– Characteristics satisfy predetermined standards

– Monitor these using externally defined measures of performance.

In-control process

– Absence of trend

– Check the stability of data-driven measures of performance.

Population

– Everything that we *might* have measured

Sample

– The part of the population that was observed

– Is it representative? (seeClass 6 of *Basic Business Statistics*)

– Aside: need to make adjustments to standard calculations if the sample represents a non-trivial fraction of the population.

Concepts

Independence

– Absence of carry-over effect from one observed value to another

– Heuristic: independence = unpredictable

– Allows us to multiply probabilities in probability calculations

Standard error of the mean

– Variation of averages from one sample to another

– Averages vary less than observations.... SD/ n

– Use in setting control limits

Central Limit Theorem

– Distribution of collection of averages grows closer to the normal as the number of underlying values being averaged increases.

– Even if individual data items are not normal, averages of such items are *approximately* normal.

Discussion

Capable versus in-control processes

– Neither feature implies the other

Why use averages in control charts?

– Averages (along with SD) completely characterize a normal population

– Detect small changes faster.

– Reduces the problem of multiplicity (fewer chances for an error)

Sampling variation

– How different might things be if I were to repeat it?

– Useful in “thought experiment”, though only conceptual as in practice we do not get to see this sort of variation.

Examples for Today

Control charts for motor shafts (page 55, 68)

– What rules should be used to monitor a production process?

– Is the process meeting standards? Is it steady over time? (Need diameter in range 810 to 820 to satisfy design.)

– Choosing what to monitor 1: individual shaft or average?
- Concern for large or small changes (see pages 78-79)
- Role for multiplicity
- Central limit theorem... averages closer to normal

– Analysis of individual shafts…
- Multiplicity and Bonferroni (page 59)
  Chance for false positive = $1 - .9973^{100} = 0.237$

– Analysis of averages of shafts
- Fewer chances for an error
  Chance for false positive = $1 - .9973^{20} = 0.053$
- Variation of averages: effects of sample size
- THE formula for standard error... $s/\sqrt{n}$ or $\sigma/\sqrt{n}$
- Control charts for means and standard deviations

– Choosing what to monitor 2: are means alone enough?

Control chart analysis of car trunk seam variation (page 80)

– Are the trunk seams of these cars meeting the design specs?

– Characteristics set by design, not incestuously from data being monitored.

– Excessive variation rather than drift in the mean values.
- Control charts for SD as well as for average.

– Alternative types of control charts (85-88)

• **Analysis of production of computer chips (Supplemental)**

– Is the output up to standards? Is it under control?

– No, this one is out of control!

A Comment

You should never see a control chart like some of these that have run so long past the point at which the mean or SD crossed the threshold. The process should have been stopped before reaching this point.