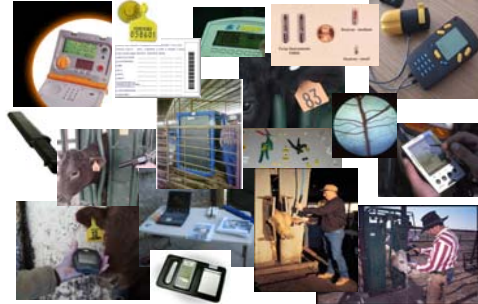




Data Collection and Analysis – Use of SPC to reveal useful information

Kevin Dhuyvetter, Ph.D.
Dept. of Ag. Economics
Kansas State University

785.532.3527
kcd@ksu.edu



Collecting and Analyzing Data

- Why should I collect data?
- Introduction to SPC and its concepts and understanding variation
- Various statistical measures and miscellaneous data issues



Why should I collect data?

- Information has value
- Potential for future use
(cannot be collected later)
- Required/regulatory
(e.g., contractor, government)



Good Business Management

- Basically all business management decisions for planning and control are based on forecasts.
- Forecasts are typically based on information (as opposed to simply being random).
- Thus, historical data are critical for making good management decisions that pertain to the future of the business.

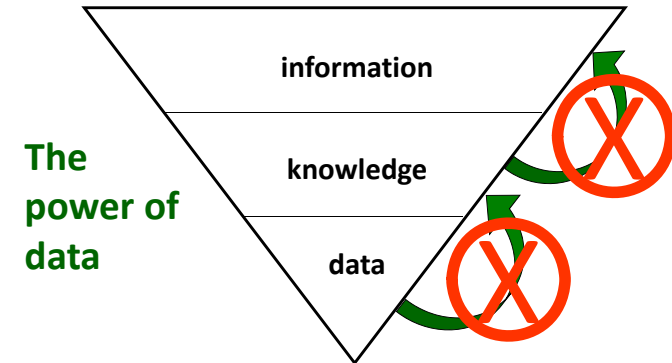


Data, knowledge, & information ...

- **Data** – facts, figures, etc., known or available; information.
- **Knowledge** – acquaintance with facts, truths, or principles, as from study or investigation; general erudition.
- **Information** – 1. knowledge communicated or received concerning some fact or circumstance; news. 2. knowledge on various subjects, of being informed.



What happens if data are wrong?



... worse yet, you don't know data are wrong and you base decisions on bad information.
(also a problem if knowledge, i.e., study of data, is incorrect)



The goal for data collection is to have useful data (*accurate* and *timely*) such that you can turn data into knowledge and information and make good business decisions.



Statistical Process Control (SPC)

The process of analyzing data so they can be transformed into knowledge and information that has value in the decision-making process.



What is SPC?

- Collection of tools, mostly statistical, which help us understand what is going on in any process that generates data, products, or services.
- Helps us attain insight into the inherent behavior of those processes which enables us to exercise control over that process.
- Serves to assist in the redesign and improvement of the process (planning & control).



Why *statistically* analyze data?

TO MAKE INFERENCES!

- Inferences: relationships expected to hold in the future
 - Optimal management: make the most profitable decisions
 - Management becomes less random



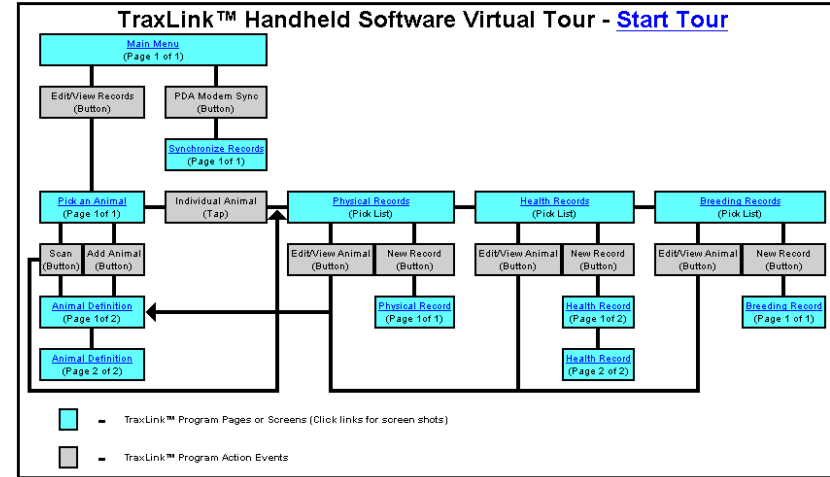
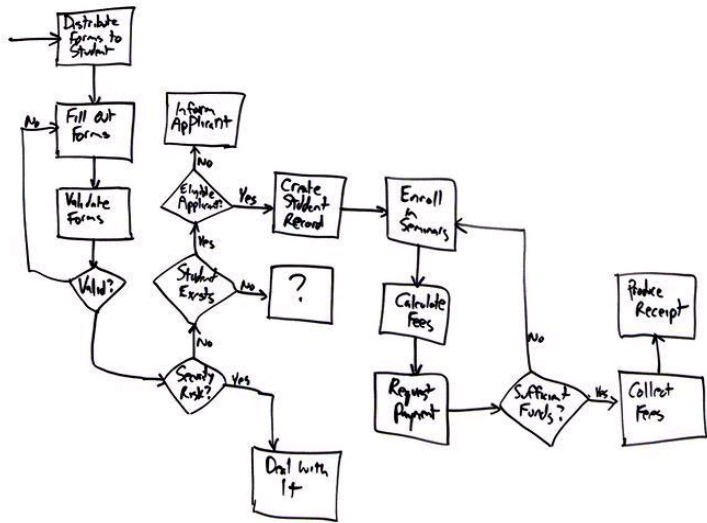
Some tools of SPC ...

- Flow charts
- Run charts
- Histograms
- Control charts
- Scatter graphs
- Summary statistics
- Sorts/queries
- Statistical analyses (regression, ANOVA)



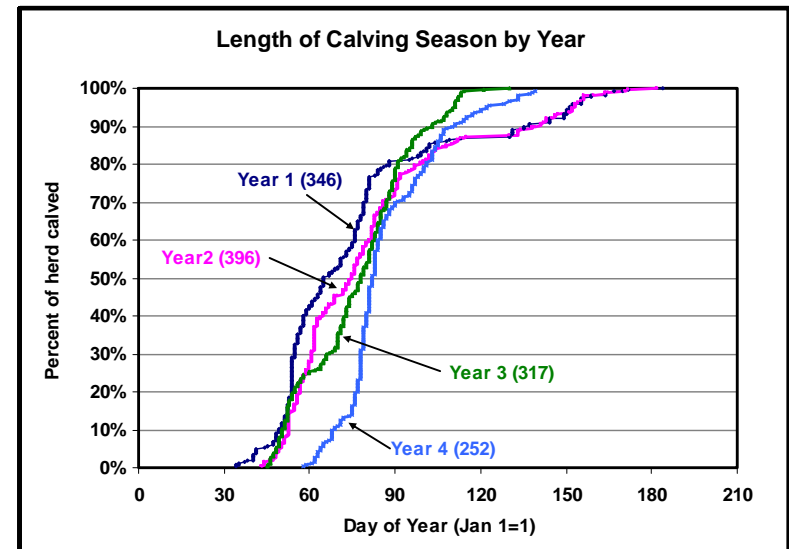
Flow charts

- A flow chart is a diagram that shows the progress of work or the flow of materials or information through a sequence of operations (also can represent an algorithm or process).
- Not considered a statistical tool, but flow charts are often useful because they provide an overall view of the entire process.



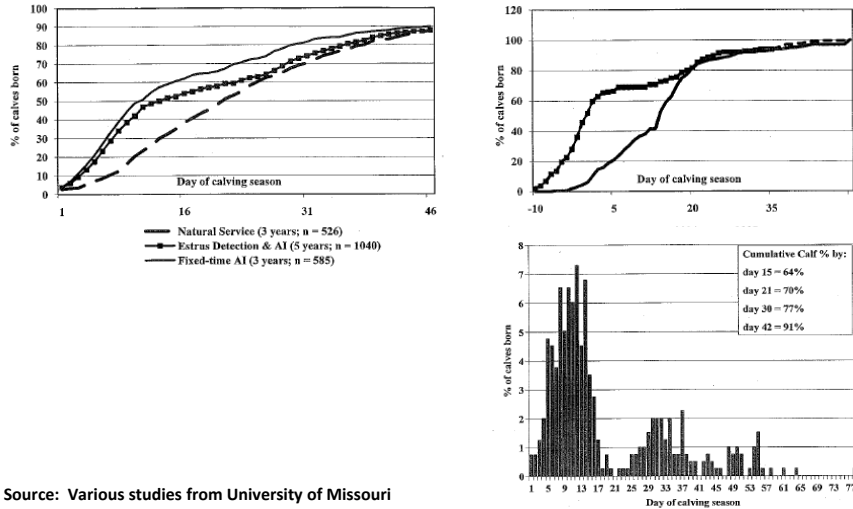
Run charts

- A run chart is a charting of sample results (data), usually one sample reading at a time, in chronological sequence.





Examples of run charts

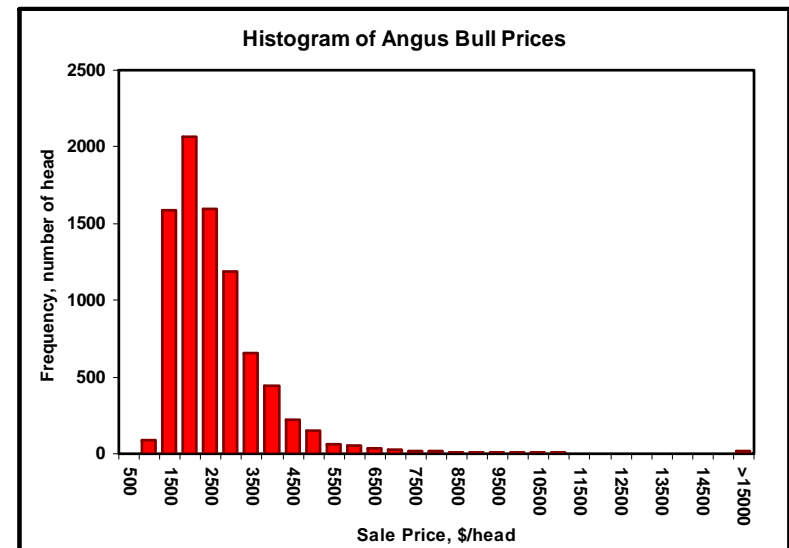
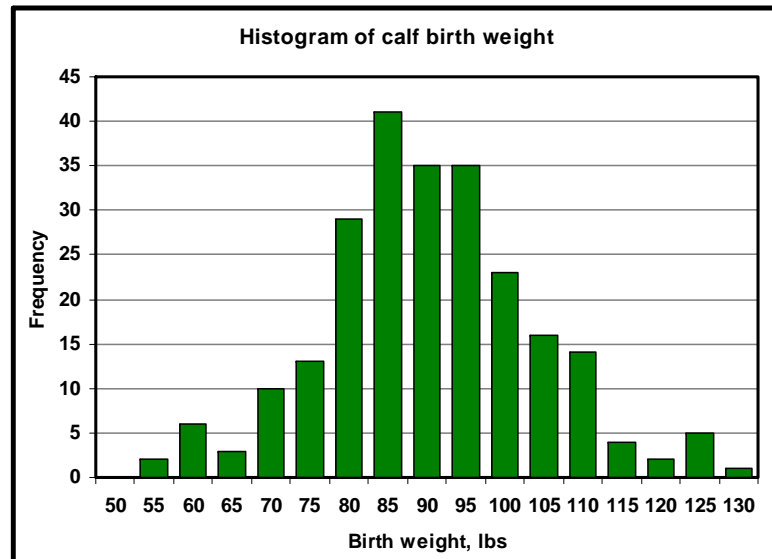


Source: Various studies from University of Missouri



Histogram (frequency distribution)

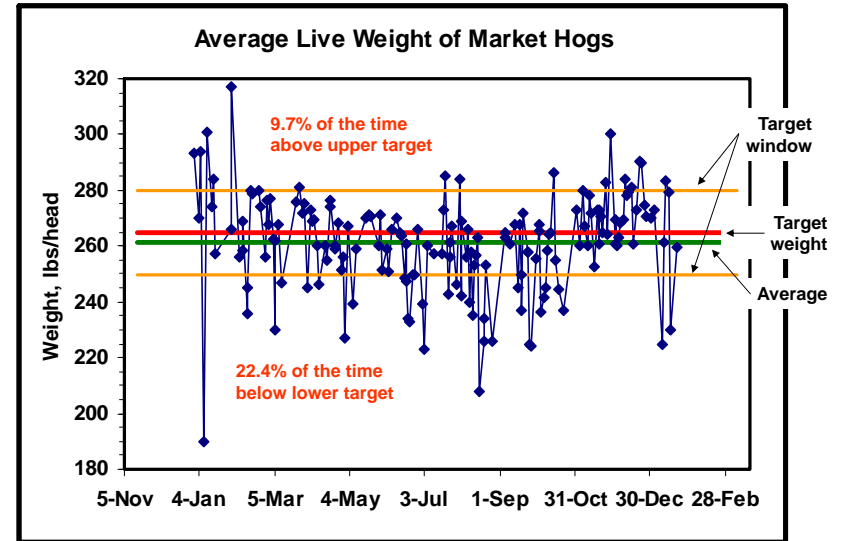
- Fundamental statistical tool of SPC
- Histograms are an effective way of summarizing data graphically that, without calculations, can illustrate three characteristics of the data.
 - A sense of the magnitude of the mean (average)
 - A sense of the variability (distribution) of the data
 - Some idea of the possible pattern of variation





Control charts

- Control charts are completely identified with SPC
- Control charts give insight into the statistical behavior of processes
- Process (def) – Everything that works together to produce a product, service or other output
 - People
 - Facilities and machinery
 - Genetics
 - Production inputs (e.g., feed, vaccines)
 - Environment

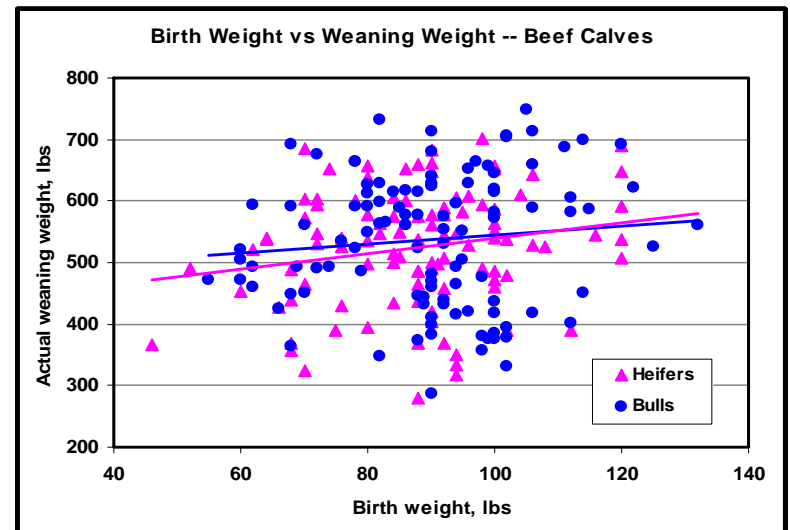


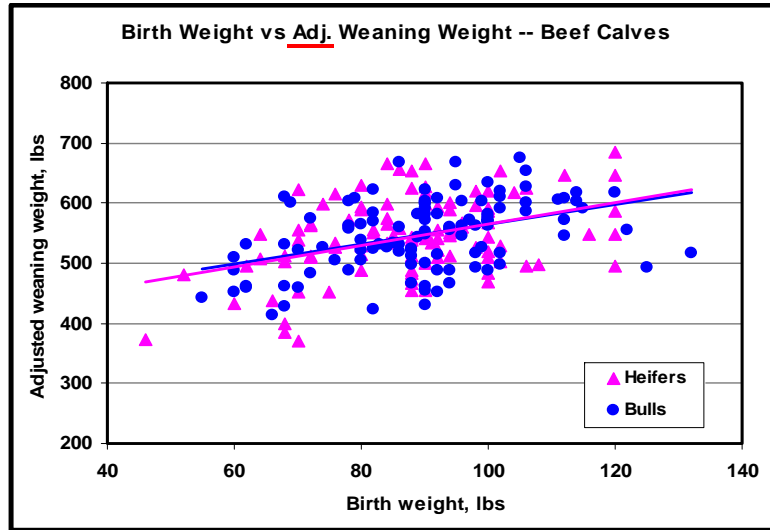
Helps identify how well you are doing, but need additional information to explain why



Scatter graphs

- Scatter diagrams help to discover associations or relationships (correlation) between two variables.
- Usually does not go beyond the graphic representation of the data (can show trend and mathematical relationship fairly easily).





Data analysis -- visual vs. numerical ...

- Visual data analysis (figures) stimulates ideas that never might emerge otherwise
- Figures get too confusing with more than one or two factors – must move to numerical analysis



Some tools of SPC...

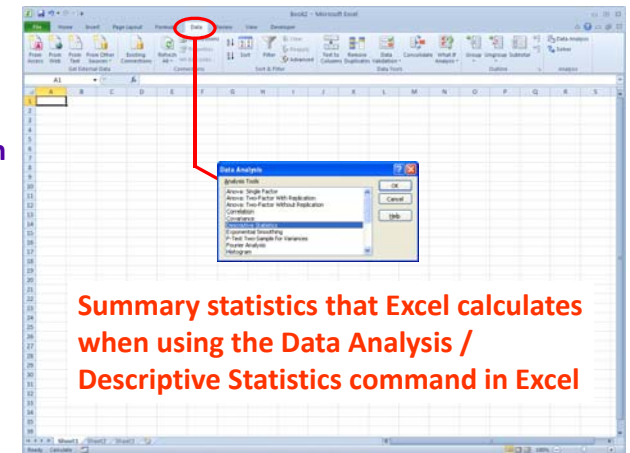
- Flow charts
- Run charts
- Histograms
- Control charts
- Scatter graphs
- Summary statistics
- Sorts/queries
- Statistical analyses

Graphical tools are extremely helpful in getting a feel for the data (i.e., first-cut analysis).

Numerical analyses are used to quantify relationships that exist for making decisions.

Summary statistics

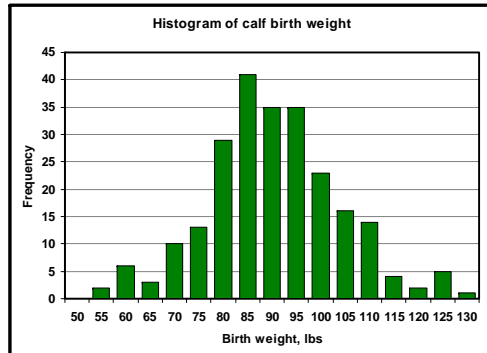
- Mean
- Standard Error
- Median
- Mode
- Standard Deviation
- Sample Variance
- Kurtosis
- Skewness
- Range
- Minimum
- Maximum
- Sum
- Count



Calf birth weight

- What is the average, how variable is it, and what kind of distribution does it follow?

Calf birth weight	
Mean	88.95
Standard Error	0.87
Median	89
Mode	80
Standard Deviation	13.45
Sample Variance	180.89
Kurtosis	0.36
Skewness	0.16
Range	74
Minimum	54
Maximum	128
Sum	21260
Count	239

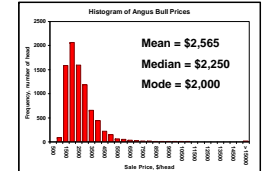


Descriptive statistics

Histogram

Measures of central tendency

- Mean – simple average (i.e., each observation is given equal weight)
- Median – middle data point in a set of values ranked from low to high
 - Median > mean (left skewed)
 - Median < mean (right skewed)
- Mode – value that occurs most frequently in a data set



Mean is often our best forecast

	farm A	farm B
Year 1	-\$80	-\$5
Year 2	\$200	\$30
Year 3	-\$50	\$20
Year 4	-\$270	\$25
Year 5	\$300	\$30
Average	\$20	\$20

How much confidence do you have in the \$20 estimate?

When analyzing data

- The mean is a powerful measure/concept
- However, the mean does not convey all important and relevant information.
- We often also want to consider the variability in the data.



Understanding variation

Causes of variation

- **Materials / inputs**
 - Raw materials exhibit variation (e.g., feed)
- **Machines**
- **Methods**
 - Ways that materials, machines and people are configured to produce goods and services
- **Genetics** (applies to both people and animals)
- **Environment**



Measures of variability

- **Range** -- the difference between the largest reading and the smallest reading.
- **Standard deviation** -- a measurement of the total variability of the data. It is an average of deviations from the mean.
- **Coefficient of variation (CV)** -- normalized measure of variability equal to standard deviation divided by the mean.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$



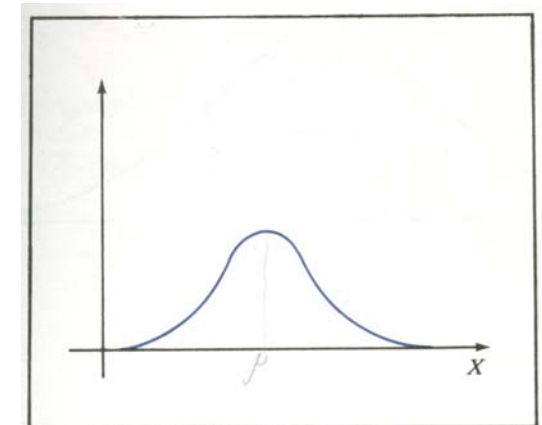
Standard deviation

- Same unit of measure as the original data
- Affected by extreme values, which tend to enlarge the standard deviation.
- Larger values for standard deviation indicate the data are more widely dispersed around the mean (i.e., more variable).
- Generally has the most meaning when data are normally distributed.

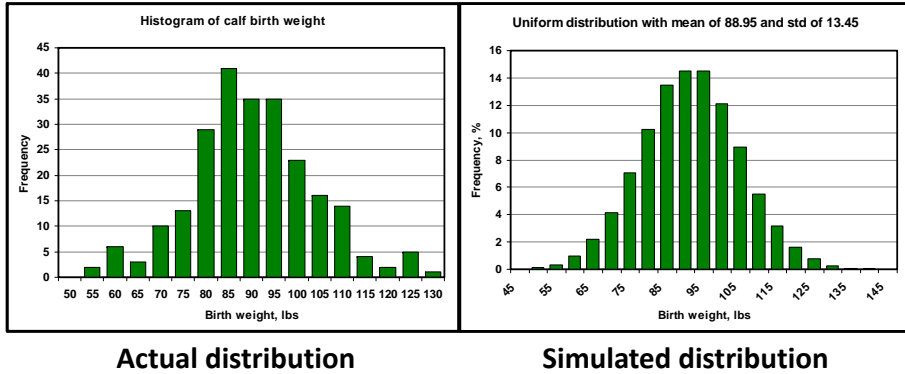


The Normal “Bell-Shaped” Curve

- **Symmetry**
- **A bell shape**
- **Smoothness**

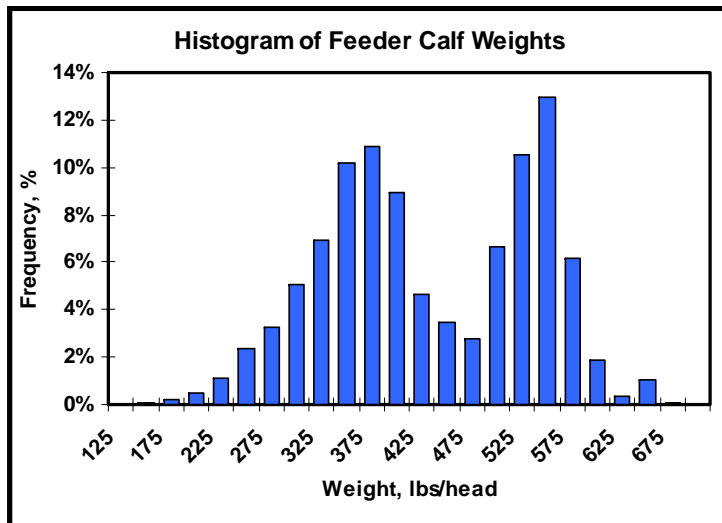
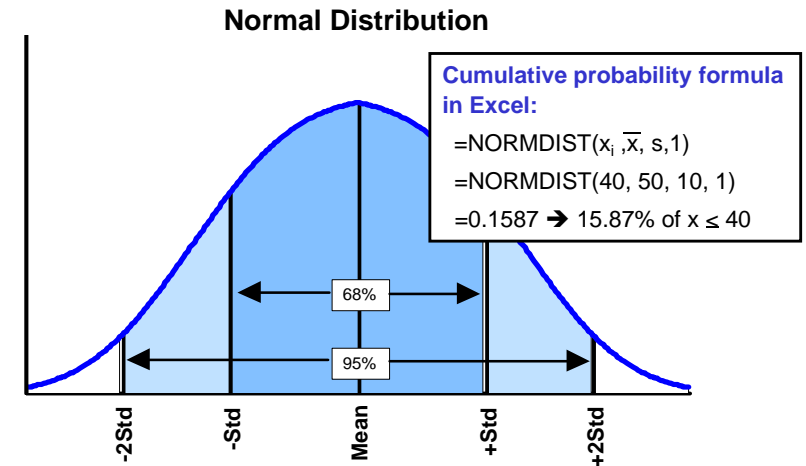


Distribution of calf birth weight

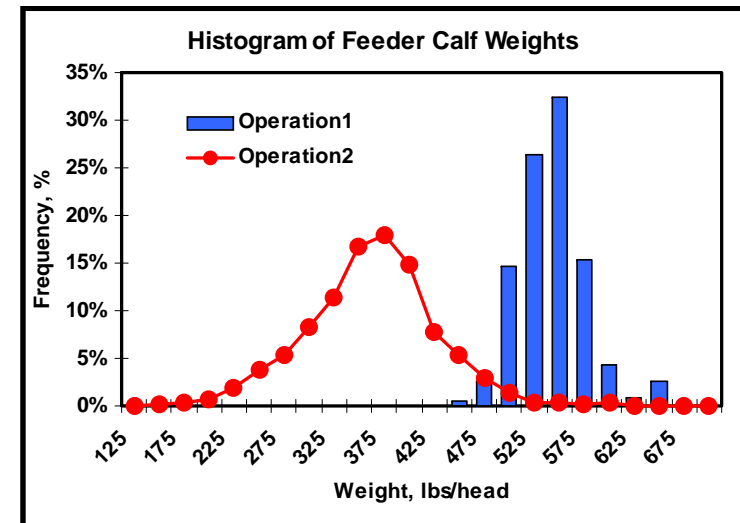


Are actual data normally distributed?

- Normal distribution -- Many statistical analysis measures assume a normal distribution



Bi-modal distribution, what might be going on?

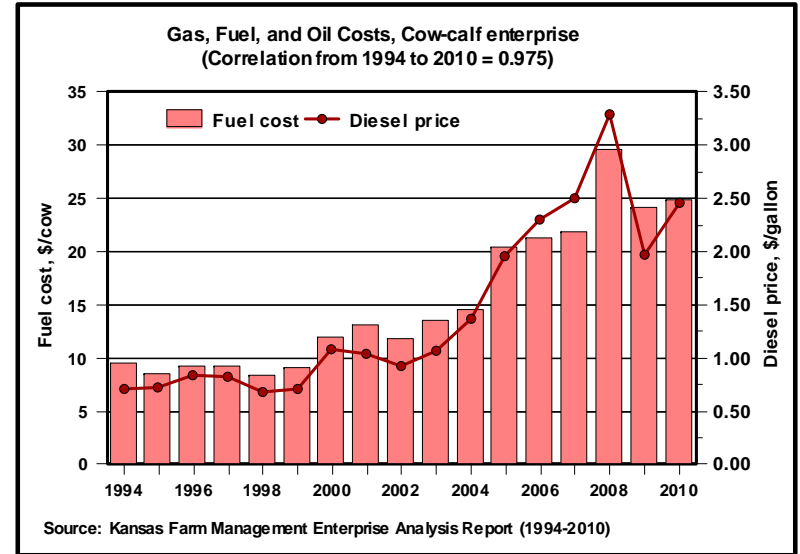


Given more information, distributions appear normal.



Statistical measures

- Correlation – degree to which two series tend to rise and fall together.
- R-square – percent variability in the dependent variable that is explained by the independent variable(s) [equal to the squared correlation between predicted and actual series].
- P-value – probability of rejecting an hypothesis when it is actually true and (1 - p-value) is an indication of the statistical confidence we can place in our results.



What does a correlation of 0.975 mean (i.e., does it have any use)?



Statistical significance

	<u>farm A</u>	<u>farm B</u>
Year 1	-\$80	-\$5
Year 2	\$200	\$30
Year 3	-\$50	\$20
Year 4	\$300	\$25
Year 5	-\$270	\$30
Average	\$20	\$20

How much confidence do you have in the \$20 estimate?

Expected value vs. confidence in expectations.

H_0 : Mean = 0; H_A : Mean > or < 0;

p-value = 0.85 (A)

p-value = 0.02 (B)



Data queries

- Conditional means, i.e., means of subgroups
 - Weaning weight (WW) for heifers
 - WW for heifers from cows 3-5 years old
 - WW for heifers from cows 3-5 years old on pasture A
 - WW for heifers from cows 3-5 years old on pasture A bred to bull X, etc...
- Requires data from many observations
- A big thing for data warehouses
- May not lead to optimal decisions



Statistical models

- Linear regression – statistical equation where variation in independent variable(s) explain variation in a dependent variable
- Generalizing “causal models”
 $y = A + Bx + e$
 $y = A + B_1x + B_2x^2 + B_3z + B_4xz + e$



Removal of outliers (bad data)

- Don't remove data believed true
- Compute mean, min, max
- Easiest way to see errors is often through visual analysis (i.e., graphically)



Normalization

- Makes it easier to compare differently scaled data
 - Weaning weights vary based on age of calf
 - Weaning weights vary based on sex of calf
 - Weaning weights vary from year to year due to environmental conditions
- Examples of normalized livestock data
 - 205-day adjusted weaning weight
 - Weight indexes (i.e., average = 100)
 - EPDs



Data analysis ...

- Preprocessing
 - Remove outliers
 - Normalization
- Statistical
 - Histograms, distributions
 - Mean, median, mode
 - Standard deviation, coefficient of variation
 - Correlation, p-value
- Data queries
- Statistical models



Why should we analyze data?

It is through the analysis of data, that “collected data” are transformed into knowledge, which then gives managers the information they need to make sound business decisions.

It's all about improving the decision-making process!



More information available at
AgManager (www.agmanager.info)

Kevin C. Dhuyvetter, Professor
Dept. of Agricultural Economics
Kansas State University
785-532-3527 -- kcd@ksu.edu

