UCLA STAT 35
Applied Computational and Interactive Probability

●**Instructor:** **Ivo Dinov,**

Asst. Prof. In Statistics and Neurology

●**Teaching Assistant:** **Anwar Khan**

University of California, Los Angeles, Winter 2005

http://www.stat.ucla.edu/~dinov/

Course Organization

Software: No specific software is required. SYSTAT, R, SOCR resource, etc.

Course Description,

Class homepage,

online supplements, VOH's, etc.

http://www.stat.ucla.edu/~dinov/courses_students.html

What is Statistics? A practical example

●*Demography: Uncertain population forecasts*

by Nico Keilman, Wolfgang Lutz, *et al.*, Nature 412, 490 - 491 (2001)

●Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But demographers and statisticians have developed methods to calculate probabilistic forecasts.

●The demographic future of any human population is uncertain, but some of the many possible trajectories are more probable than others. So, forecast demographics of a population, e.g., size by 2100, should include two elements: a range of possible outcomes, and a probability attached to that range.

What is Statistics?

●Together, ranges/probabilities constitute a *prediction interval* for the population. There are trade-offs between greater certainty (higher odds) and better precision (narrower intervals). Why?

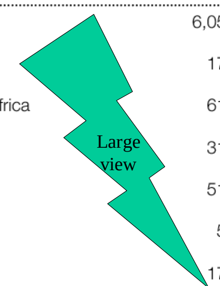
●For instance, the next table shows an estimate that the odds are 4 to 1 (an 80% chance) that the world's population, now at 6.1 billion, will be in the range [5.6 : 12.1] billion in the year 2100. Odds of 19 to 1 (a 95% chance) result in a **wider** interval: [4.3 : 14.4] billion.

Table 1 Forecasted population sizes and proportions over age 60

	Median world and regional population sizes (millions)				
Year	2000	2025	2050	2075	2100
World total	6,055	7,827 (7,219–8,459)	8,797 (7,347–10,443)	8,951 (6,636–11,652)	8,414 (5,577–12,123
North Africa	173	257 (228–285)	311 (249–378)	336 (238–443)	333 (215–484)
Sub-Saharan Africa	611	976 (856–1,100)	1,319 (1,010–1,701)	1,522 (1,021–2,194)	1,500 (878–2,450)
North America	314	379 (351–410)	422 (358–498)	441 (343–565)	454 (313–631)
Latin America	515	709 (643–775)	840 (679–1,005)	904 (647–1,202)	934 (585–1,383)
Central Asia	56	81 (73–90)	100 (80–121)	107 (76–145)	106 (66–159)
Middle East	172	285 (252–318)	368 (301–445)	413 (296–544)	413 (259–597)
South Asia	1,367	1,940 (1,735–2,154)	2,249 (1,795–2,776)	2,242 (1,528–3,085)	1,958 (1,186–3,035)
China region	1,408	1,608 (1,494–1,714)	1,580 (1,305–1,849)	1,422 (1,003–1,884)	1,250 (765–1,870)
Pacific Asia	476	625 (569–682)	702 (575–842)	702 (509–937)	654 (410–949)
Pacific OECD	150	155 (144–165)	148 (125–174)	135 (100–175)	123 (79–173)
Western Europe	456	478 (445–508)	470 (399–549)	433 (321–562)	392 (257–568)
Eastern Europe	121	117 (109–125)	104 (86–124)	87 (61–118)	74 (44–115)
European part of the former USSR	236	218 (203–234)	187 (154–225)	159 (110–216)	141 (85–218)

Table 1 Forecasted population sizes and proportions over

		Median
Year	2000	2025
World total	6,055	7,827 (7,219–8,459)
North Africa	173	257 (228–285)
Sub-Saharan Africa	611	976 (856–1,100)
North America	314	379 (351–410)
Latin America	515	709 (643–775)
Central Asia	56	81 (73–90)
Middle East	172	285 (252–318)



What is Statistics?

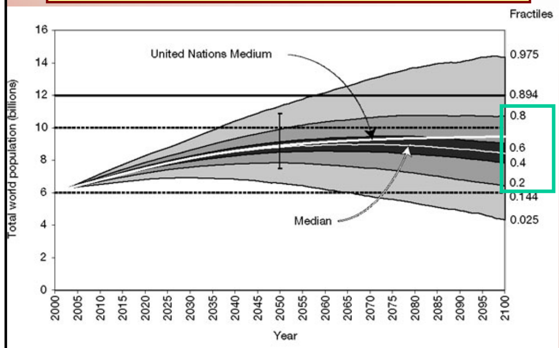
- *Demography: Uncertain population forecasts*

  by Nico Keilman, Nature 412, ,2001
- Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But lately demographers and statisticians have developed methods to calculate probabilistic forecasts.
- Proportion of population over 60yrs.

Proportion of population over age 60

2000	2050	2100
0.10	0.22 (0.18–0.27)	0.34 (0.25–0.44)
0.06	0.19 (0.15–0.25)	0.32 (0.23–0.44)
0.05	0.07 (0.05–0.09)	0.20 (0.14–0.27)
0.16	0.30 (0.23–0.37)	0.40 (0.28–0.52)
0.08	0.22 (0.17–0.28)	0.33 (0.23–0.45)
0.08	0.20 (0.15–0.25)	0.34 (0.24–0.46)
0.06	0.18 (0.14–0.23)	0.35 (0.24–0.47)
0.07	0.18 (0.14–0.24)	0.35 (0.25–0.48)
0.10	0.30 (0.24–0.37)	0.39 (0.27–0.53)
0.08	0.23 (0.18–0.29)	0.36 (0.26–0.49)
0.22	0.39 (0.32–0.47)	0.49 (0.35–0.61)
0.20	0.35 (0.29–0.43)	0.45 (0.32–0.58)
0.18	0.38 (0.30–0.46)	0.42 (0.28–0.57)
0.19	0.35 (0.27–0.44)	0.36 (0.23–0.50)

What is Statistics?



What is Statistics?

- There is concern about the accuracy of population forecasts, in part because the <u>rapid fall in fertility in Western countries in the 1970s</u> came as a surprise. Forecasts made in those years predicted birth rates that were up to 80% too high.
- The rapid reduction in mortality after the Second World War was also not foreseen; life-expectancy forecasts were too low by 1–2 years; and the predicted number of elderly, particularly the oldest people, was far too low.

What is Statistics?

- So, during the 1990s, researchers developed methods for making probabilistic population forecasts, the **aim** of which is to calculate prediction intervals for every variable of interest. Examples include population forecasts for the USA, AU, DE, FIN and the Netherlands; these forecasts comprised prediction intervals for <u>variables</u> such as age structure, average number of children per woman, immigration flow, disease epidemics.
- We need accurate probabilistic population forecasts for the whole world, and its 13 large division regions (see Table). The <u>conclusion</u> is that there is an estimated 85% chance that the world's population will stop growing before 2100. Accurate?

What is Statistics?

- There are <u>three main methods of probabilistic forecasting</u>:
  - time-series extrapolation;
  - expert judgement; and
  - extrapolation of historical forecast errors.
- Time-series methods rely on statistical models that are fitted to historical data. These methods, however, seldom give an accurate description of the past. If many of the historical facts remain unexplained, time-series methods result in excessively wide prediction intervals when used for <u>long-term forecasting</u>.
- Expert judgement is subjective, and historic-extrapolation alone may be near-sighted.

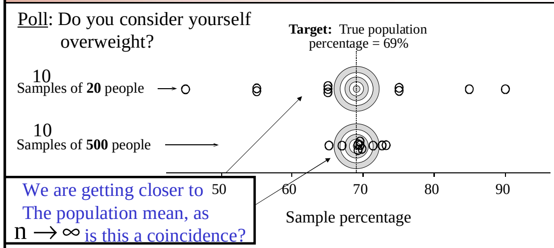
Intro & Descriptive Stats

- Variation in data
- Data Distributions
- Stationary and (dynamic) non-stationary processes
- Causes of Variation

Newtonian science vs. chaotic science

- Article by Robert May, Nature, vol. 411, June 21, 2001
  - Science we encounter at schools deals with crisp certainties (e.g., prediction of planetary orbits, the periodic table as a descriptor of all elements, equations describing area, volume, velocity, position, etc.)
  - As soon as uncertainty comes in the picture it shakes the foundation of the deterministic science, because only probabilistic statements can be made in describing a phenomenon (e.g., roulette wheels, chaotic dynamic weather predictions, Geiger counter, earthquakes, etc.)
  - What is then science all about – describing absolutely certain events and laws alone, or describing more general phenomena in terms of their behavior and chance of occurring? Or may be both!

Variation in sample percentages



Comparing percentages from 10 different surveys each of 20 people with those from 10 surveys each of 500 people (all surveys from same population).

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

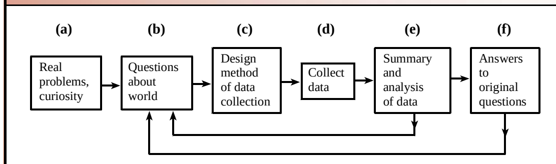
Experiments vs. observational studies for comparing the effects of treatments

- In an Experiment
  - experimenter determines which units receive which treatments. (ideally using some form of random allocation)
- Observational study – useful when can't design a controlled randomized study
  - compare units that happen to have received each of the treatments
  - Ideal for describing relationships between different characteristics in a population.
  - often useful for identifying possible causes of effects, but cannot reliably establish causation.
- Only properly designed and executed experiments can reliably demonstrate causation.

The Subject of Statistics

- Statistics is concerned with the process of finding out about the world and how it operates - in the face of variation and uncertainty
- by collecting and analyzing, making sense (interpreting) of data.
- Data are measurements, facts and information about an object or a process that allows is to make inference about the object being observed.

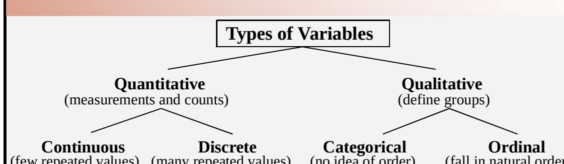
The investigative process



The investigative process.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Distinguishing between types of variable



Experimental vs. Observation study

- A researcher wants to evaluate IQ levels are related to person's height. <u>100 people</u> are are randomly selected and grouped into <u>5 bins</u>: [0:50), [50;100), [100:150], [150:200), [200:250] *cm* in height. The subjects undertook a IQ exam and the results are analyzed.
- Another researcher wants to assess the bleaching effects of <u>10 laundry detergents</u> on <u>3 different colors</u> (R,G,B). The laundry detergents are randomly selected and applied to 10 pieces of cloth. The discoloration is finally evaluated.

Experimental vs. Observation study

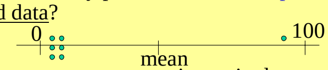
- For each study, describe what *treatment* is being compared and what *response* is being measured to compare the treatments.
- Which of the studies would be described as *experiments* and which would be described as *observational* studies?
- For the studies that are observational, could an experiment have been carried out instead? If not, briefly explain why not.
- For the studies that are experiments, briefly discuss what *forms of blinding* would be possible to be used.
- In which of the studies has *blocking* been used? Briefly describe *what* was blocked and why it was blocked.

Experimental vs. Observation study

- What is the *treatment* and what is the *response*?
  1. <u>Treatment</u> is height (as a bin). <u>Response</u> is IQ score.
  2. <u>Treatment</u> is laundry detergent. <u>Response</u> is discoloration.
- *Experiment* or *observational* study?
  1. Observational – compare obs's (IQ) which happen to have the treatment (height).
  2. Experimental – experimenter controls which treatment is applied to which unit.
- For the <u>observational</u> studies, can we conduct an experiment?
  1. This could not be done as an experiment - it would require the experimenter to decide the (natural) height (treatment) of the subjects (units).
- For the <u>experiments</u>, is there *blinding?*
  2. The only form of blinding possible would be for the technicians measuring the cloth discoloration not to know which detergent was applied.
- Is there *blocking*?
  1. & 2. No blocking. Say, if there are two laundry machines with different cycles of operation and if we want to block we'll need to randomize which laundry does which cloth/detergent combinations, because differences in laundry cycles are a known source of variation.

Mean, Median, Mode, Quartiles, 5# summary

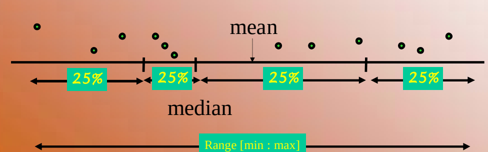
- The sample mean is the average of all numeric obs's.
- The sample median is the obs. at the index $(n+1)/2$ (note take avg of the 2 obs's in the middle for fractions like 23.5), of the observations ordered by size (small-to-large)?
- The sample median usually preferred to the sample mean for <u>skewed data</u>?



- Under what circumstances may quoting a <u>single center</u> (be it mean or median) not make sense?(multi-modal)
- What can we say about the sample mean of a qualitative variable? (meaningless)

Quartiles

The first quartile (Q_1) is the median of all the observations whose *position* is <u>strictly below the *position* of the median</u>, and the third quartile (Q_3) is the median of those above.



Five number summary

The five-number summery = (Min, Q_1, Med, Q_3, Max)

Quantiles (vs. quartiles)

- The q^{th} **quantile** (100 x q^{th} percentile) is a value, in the range of our data, so that proportion of at least **q** of the data lies at or below it and a proportion of at least **(1-q)** lies at or above it.
- E.x., X={1,2,3,4,5,6,7,8,9,10}. The **20th** percentile (0.2 quartile) is the value **2**, since 20% of the data is below it and 80% above it. The **70th** percentile is the value 7, etc.
- We could have also selected **2.5** and **7.5** for the **20th** and **70th** percentile , above. There is no agreement on the exact definitions of quantiles.

Measures of variability (deviation)

- **Mean Absolute Deviation (MAD)** –

$$MAD = \frac{1}{n-1}\sum_{i=1}^{n}\left|y_i - \bar{y}\right|$$

- **Variance** –

$$Var = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

- **Standard Deviation** –

$$SD = \sqrt{Var} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Measures of variability (deviation)

- **Example:**
- **Mean Absolute Deviation**– $MAD = \frac{1}{n-1}\sum_{i=1}^{n}\left|y_i - \bar{y}\right|$
- **Variance** – $Var = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$
- **Standard Deviation** – $SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$
- X={1, 2, 3, 4}.

m=2.5

1 2 3 4

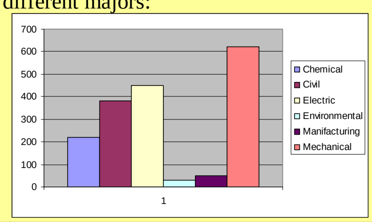
MAD=4/3=1.33

Var=5/3=1.67

SD=1.3

Bar Chart

- List all possible categories the data is classified in!
  - Represents the frequency of occurrence of the data in each category
- Example: Number of engineering students enrolled in different majors:



220

380

450

30

50

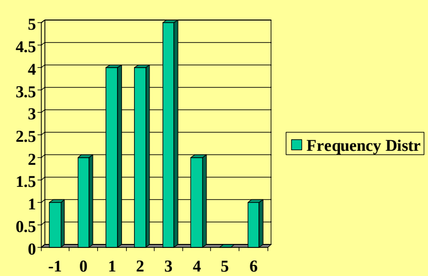
620

Data Distribution

- A data distribution is a summary of the variation in a dataset. Data distribution is a list of all possible values (of the process/object) and their respective frequencies (e.g., how often is each possible value encountered, when we observe the object/process).
- E.g., {**1, 2, 2, 3, -1, 0, 0, 1, 2, 3, 4, 3, 3, 1, 2, 1, 4, 6, 3**}

Data Distribution

- E.g., {**1, 2, 2, 3, -1, 0, 0, 1, 2, 3, 4, 3, 3, 1, 2, 1, 4, 6, 3**}

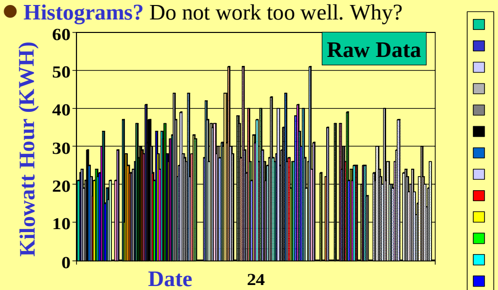


Stationarity of Processes

- Does the variability of the data change significantly as more data is collected (say between different time points, different physical locations, etc.)?
- **Stationary process** is a data-generating mechanism for which the distribution of the resulting data does NOT change appreciably as more data is being observed.
- **Non-Stationary process** is a data-generating mechanism for which the distribution of the resulting data DOES change as more data is being observed.
- E.g., Grades (over time), Air quality (in different regions in the US), Geiger counter (time), Species Extinction (long-times). Other examples?

Stationary or Non-Stationary Process?

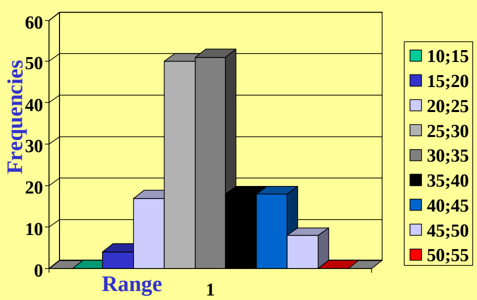
- **Histograms?** Do not work too well. Why?



Data: (elece1.xls) 174 daily obs’s, one household

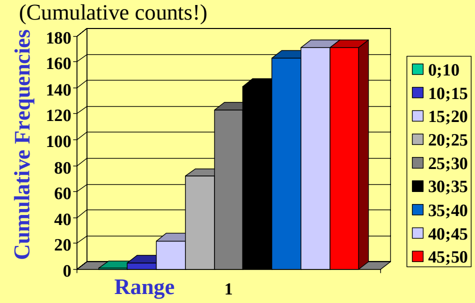
Stationary or Non-Stationary Process?

- **Histograms?** Do not work too well. Why?



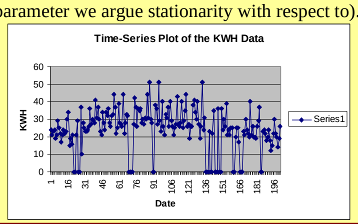
Stationary or Non-Stationary Process?

- **Histograms?** Do not work too well. Why? (Cumulative counts!)



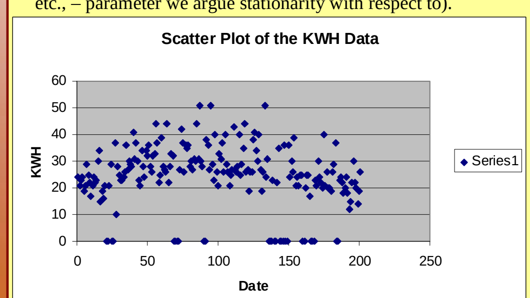
Stationary or Non-Stationary Process?

- **To assess stationarity:**
- **Rigorous assessment:** A stationary process has a **constant mean**, **variance**, and **autocorrelation** through time/space.
- **Visual assessment:** (Plot the data – observed vs. time/place – the parameter we argue stationarity with respect to).



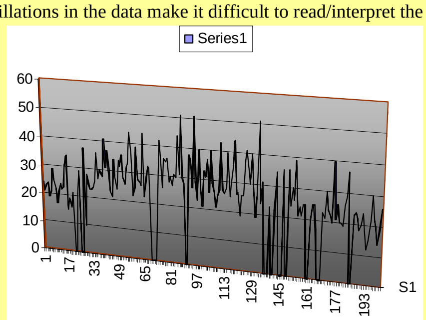
Stationary or Non-Stationary Process?

- **Visual assessment:** (Plot the data – observed vs. time/place, etc., – parameter we argue stationarity with respect to).



Moving Averages

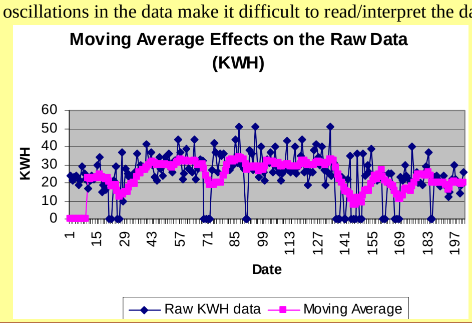
- **Signal, Noise, Filtering:** Oftentimes high frequency oscillations in the data make it difficult to read/interpret the data.



Slide 42 Stat 35, UCLA, Ivo Dinov

Moving Averages – next 10 values are averaged

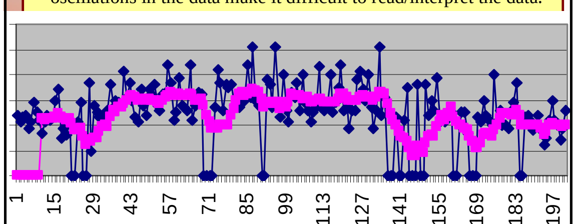
- **Signal, Noise, Filtering:** Oftentimes high frequency oscillations in the data make it difficult to read/interpret the data.



Slide 43 Stat 35, UCLA, Ivo Dinov

Moving Averages – zoom in

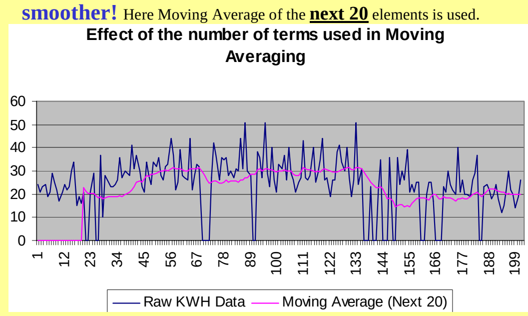
- **Signal, Noise, Filtering:** Oftentimes high frequency oscillations in the data make it difficult to read/interpret the data.



Slide 44 Stat 35, UCLA, Ivo Dinov

Moving Averages - next 20

- **Averages with more terms being averaged are smoother!** Here Moving Average of the **next 20** elements is used.



Slide 45 Stat 35, UCLA, Ivo Dinov

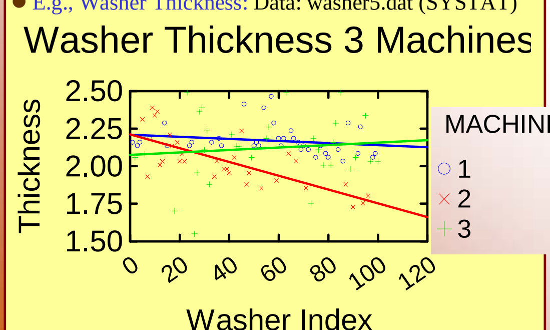
Causes of Variation in Data

- Cause of variation is the reason/mechanism that introduces some of the observed variation in the data.
- Kinds of causes of variation:
  - Common cause – the inherited fluctuations in a process, e.g., Geiger counter variances, random arrival time variances
  - Special causes – periodically/cyclically arising variances, e.g., temp measures vary with season, wake-up times vary specially with day-of-week (weekends most people sleep longer), different machine settings/protocols (MRI imaging).

Slide 46 Stat 35, UCLA, Ivo Dinov

Causes of Variation in Data

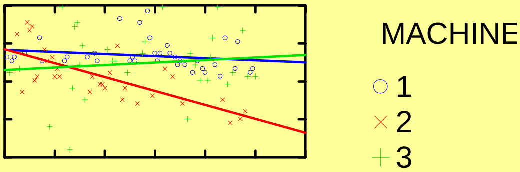
- E.g., Washer Thickness: Data: washer5.dat (SYSTAT)



Slide 47 Stat 35, UCLA, Ivo Dinov

Causes of Variation in Data – Stratified Plots for identifying causes of variation

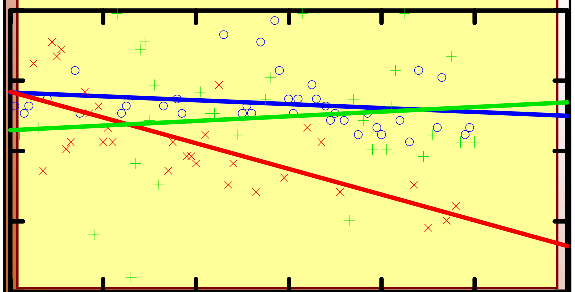
- E.g., Washer Thickness: Data: washer5.dat (SYSTAT)



- Machine 1 exhibits normal operating conditions (variation purely due to common causes). Machine 3 shows abnormal cyclic variation. Machine 2 shows a consistent robust thickness decline – special causes variability.

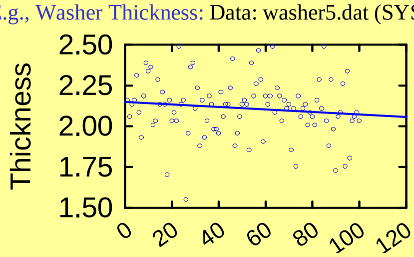
Causes of Variation in Data – Stratified Plots for identifying causes of variation

- E.g., Washer Thickness: Data: washer5.dat (SYSTAT)



Causes of Variation in Data

- E.g., Washer Thickness: Data: washer5.dat (SYSTAT)



- If you plot all washer thicknesses all together you may miss the facts that there are special cause effects on the variances.

Trimmed, Winsorized means and Resistancy

- A data-driven **parameter estimate** is said to be **resistant** if it does not greatly change in the presence of outliers.
- **K-times trimmed mean**

$$\bar{y}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} y_{(i)}$$

Order statistic

- **Winsorized k-times mean**:

$$\bar{y}_{wk} = \frac{1}{n}\left[(k+1)y_{(k+1)} + \sum_{i=k+2}^{n-k-1} y_{(i)} + (k+1)y_{(n-k)}\right]$$

Example - Trimmed, Winsorized means and Resistancy

- **K-times trimmed mean** $\bar{y}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} y_{(i)}$
- **Winsorized k-times mean**:

$$\bar{y}_{wk} = \frac{1}{n}\left[(k+1)y_{(k+1)} + \sum_{i=k+2}^{n-k-1} y_{(i)} + (k+1)y_{(n-k)}\right]$$

- Data: {-11, 2, -1, 0, 1, 2, 0, -1,15, 100}, n=10, **Say k=2**
- Ordered statistics $y_{(i)}$: {-11, -1, -1, 0,0,1, 2, 2, 15, 100}

$$\bar{y} = \frac{1}{10}[-11-1+...+15+100] = 107/10 \sim 11$$

$$\bar{y}_{tk} = \frac{1}{10-4}(-1+0+0+1+2+2) = 4/6$$

$$\bar{y}_{wk} = \frac{1}{10}[3(-1)+(0+0+1+2)+3\times 2] = 3/5$$