

UCLA STAT 35
Applied Computational and Interactive Probability

● **Instructor:** Ivo Dinov,
 Asst. Prof. In Statistics and Neurology

● **Teaching Assistant:** Chris Barr

University of California, Los Angeles, Winter 2006
<http://www.stat.ucla.edu/~dinov/>

Stat 35, UCLA, Ivo Dinov Slide 1

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

Slide 2 Stat 35, UCLA, Ivo Dinov

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.

Stat 35, UCLA, Ivo Dinov Slide 3

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.

Stat 35, UCLA, Ivo Dinov Slide 4

Table 1 Forecasted population sizes and proportions over age 60

Year	Median world and regional population sizes (millions)			
	2000	2025	2050	2075
World total	6,055	7,827	8,797	8,951
North Africa	173	257	311	336
Sub-Saharan Africa	611	(228-285)	(249-378)	(238-443)
North America	314	(856-1,100)	(1,010-1,701)	(1,021-2,194)
Latin America	515	709	840	904
Central Asia	56	81	100	107
Middle East	172	(73-90)	(80-121)	(76-145)
South Asia	1,367	1,940	2,249	2,242
China region	1,408	(1,735-2,154)	(1,795-2,778)	(1,328-3,085)
Pacific Asia	476	525	702	702
Pacific OECD	150	(569-682)	(575-842)	(509-907)
Western Europe	456	478	470	433
Eastern Europe	121	(117)	104	87
European part of the former USSR	236	(109-125)	(86-124)	(61-118)
		218	159	141
		(203-234)	(154-225)	(110-216)

Stat 35, UCLA, Ivo Dinov Slide 5

Table 1 Forecasted population sizes and proportions over age 60

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

Large view

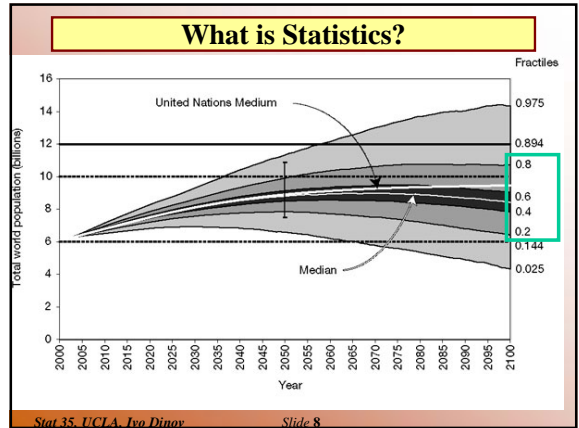
Stat 35, UCLA, Ivo Dinov Slide 5

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

Stat 35, UCLA, Ivo Dinov Slide 7



What is Statistics?

- There is concern about the **accuracy of population forecasts**, in part because the **rapid fall in fertility in Western countries in the 1970s** 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**.

Stat 35, UCLA, Ivo Dinov Slide 9

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 **variables** 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 **conclusion** is that there is an estimated 85% chance that the **world's population will stop growing before 2100**. Accurate?

Stat 35, UCLA, Ivo Dinov Slide 10

What is Statistics?

- There are **three main methods of probabilistic forecasting**:
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 **long-term forecasting**.
- **Expert judgement** is subjective, and **historic-extrapolation** alone may be near-sighted.

Stat 35, UCLA, Ivo Dinov Slide 11

Intro & Descriptive Stats

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

Stat 35, UCLA, Ivo Dinov Slide 12

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.)

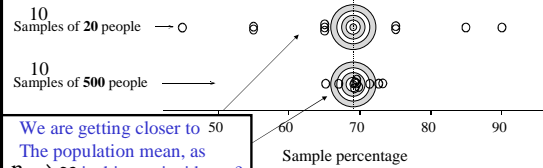
● As soon as **uncertainty** comes in the picture it **shakes the foundation of the deterministic science**, because only **probabilistic statements** can be made in describing a phenomenon (e.g., roulette wheels, chaotic dynamic weather predictions, Geiger counter, earthquakes, etc.)

● **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!

Slide 13 Stat 35, UCLA, Jon Dinger

Variation in sample percentages

Poll: Do you consider yourself overweight?



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.

Slide 14 Stat 35, UCLA, Jon Dinger

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**.

Slide 15 Stat 35, UCLA, Jon Dinger

The Subject of Statistics

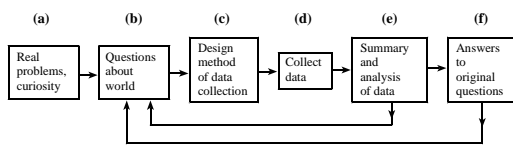
● **Statistics** is concerned with the process of finding out about the world and how it operates - in the face of **variation** and **uncertainty**

● 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.

Slide 16 Stat 35, UCLA, Jon Dinger

The investigative process

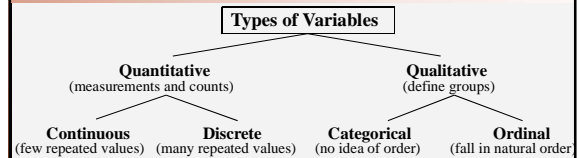


The investigative process.

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

Slide 17 Stat 35, UCLA, Jon Dinger

Distinguishing between types of variable



Slide 23 Stat 35, UCLA, Jon Dinger

Experimental vs. Observation study

- A researcher wants to evaluate IQ levels are related to person's height. 100 people are randomly selected and grouped into 5 bins: [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 10 laundry detergents on 3 different colors (R,G,B). The laundry detergents are randomly selected and applied to 10 pieces of cloth. The discoloration is finally evaluated.

Slide 24 Stat 35, UCLA, Jon Dinger

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.

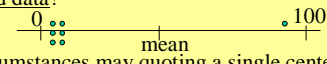
Slide 25 Stat 35, UCLA, Jon Dinger

Experimental vs. Observation study

- What is the *treatment* and what is the *response*?
 1. Treatment is height (as a bin). Response is IQ score.
 2. Treatment is laundry detergent. Response 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 *observational* 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 *experiments*, 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.

Slide 26 Stat 35, UCLA, Jon Dinger

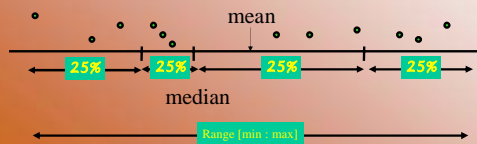
Mean, Median, Mode, Quartiles, 5# summary

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

Slide 27 Stat 35, UCLA, Jon Dinger

Quartiles

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



Slide 28 Stat 35, UCLA, Jon Dinger

Five number summary

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

Slide 29 Stat 35, UCLA, Jon Dinger

Quantiles (vs. quartiles)

- The **qth quantile** (100 x **qth 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.

Slide 30 Stat 35, UCLA, Jon Dinger

Measures of variability (deviation)

- **Mean Absolute Deviation (MAD)** –

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

- **Variance** –

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

- **Standard Deviation** –

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

Slide 31 Stat 35, UCLA, Jon Dinger

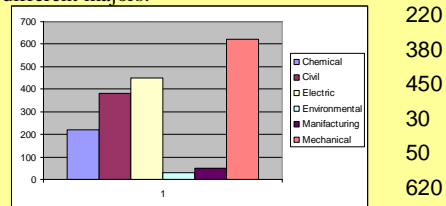
Measures of variability (deviation)

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

Slide 32 Stat 35, UCLA, Jon Dinger

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:



Slide 33 Stat 35, UCLA, Jon Dinger

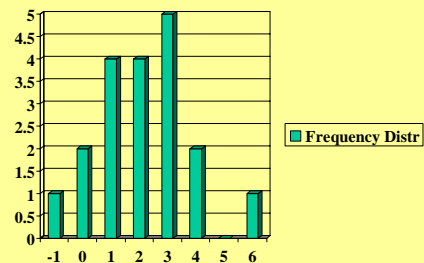
Data Distribution

- 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}

Slide 34 Stat 35, UCLA, Jon Dinger

Data Distribution

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



Slide 35 Stat 35, UCLA, Jon Dinger

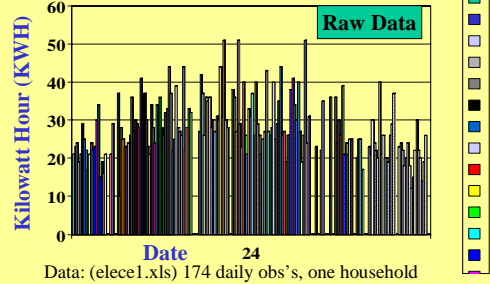
Stationarity of Processes

- 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?

Slide 36 Stat 35, UCLA, Jon Dinger

Stationary or Non-Stationary Process?

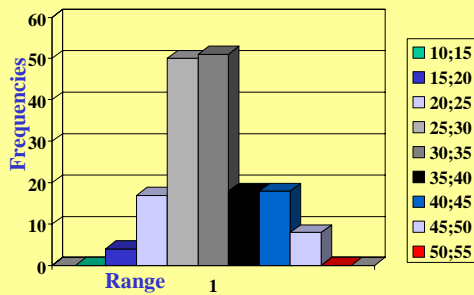
- Histograms?** Do not work too well. Why?



Slide 37 Stat 35, UCLA, Jon Dinger

Stationary or Non-Stationary Process?

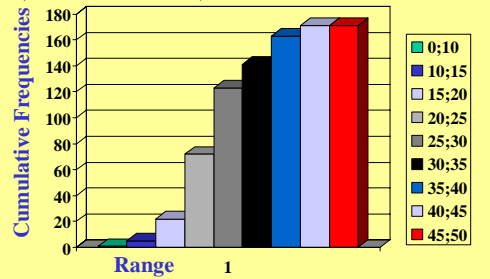
- Histograms?** Do not work too well. Why?



Slide 38 Stat 35, UCLA, Jon Dinger

Stationary or Non-Stationary Process?

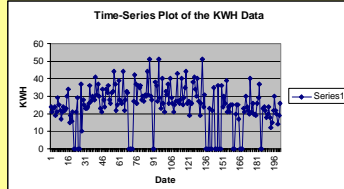
- Histograms?** Do not work too well. Why? (Cumulative counts!)



Slide 39 Stat 35, UCLA, Jon Dinger

Stationary or Non-Stationary Process?

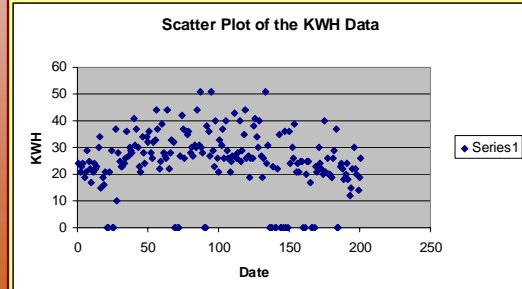
- To assess stationarity:**
- Rigorous assessment:** A stationary process has a **constant mean, variance, and autocorrelation** through time/space.
- Visual assessment:** (Plot the data – observed vs. time/place – the parameter we argue stationarity with respect to).



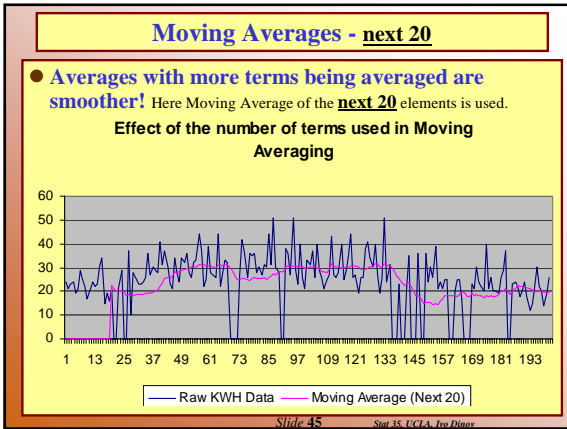
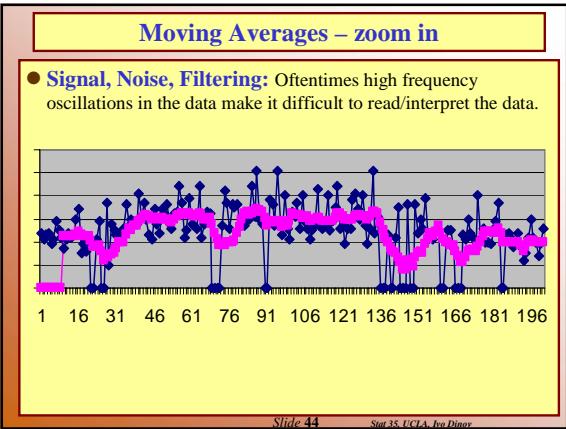
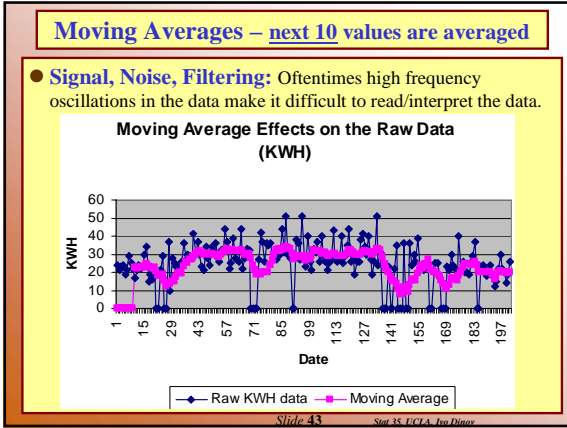
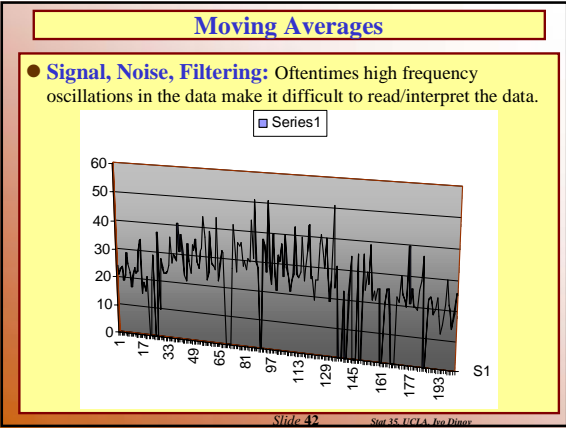
Slide 40 Stat 35, UCLA, Jon Dinger

Stationary or Non-Stationary Process?

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



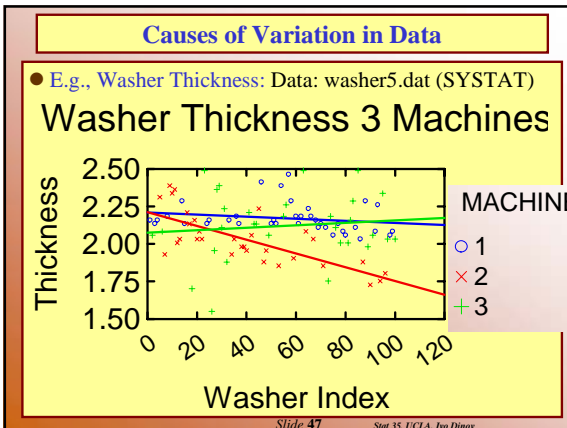
Slide 41 Stat 35, UCLA, Jon Dinger

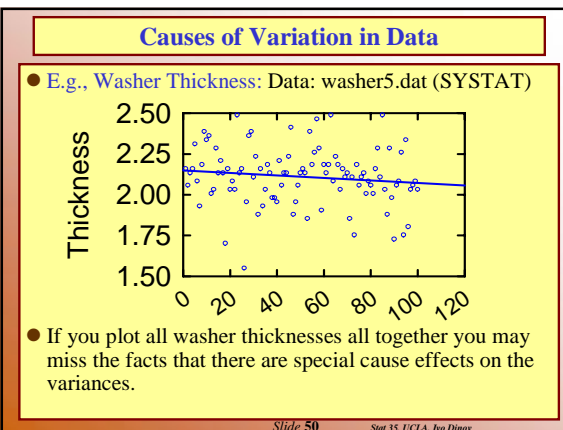
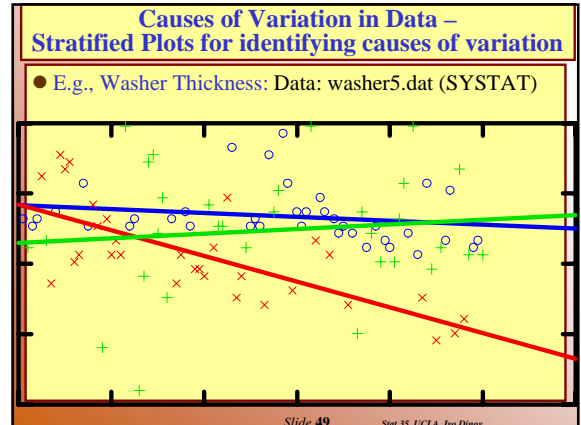
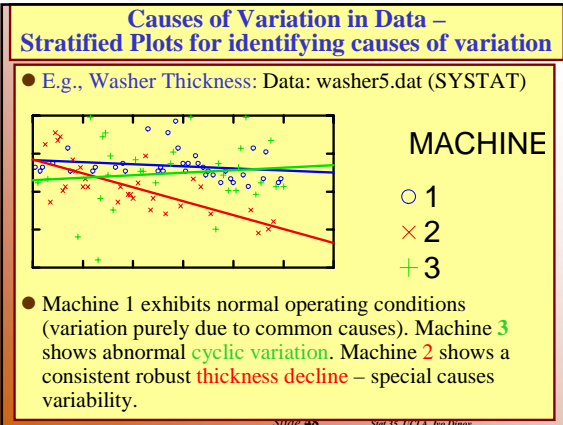


Causes of Variation in Data

- **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, Jon Dinger





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**

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

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

Slide 51 Stat 35, UCLA, Jon Dinger

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}
- $\bar{y} = \frac{1}{10}[-11-1+\dots+15+100] = 107/10 \sim 11$
- $\bar{y}_{tk} = \frac{1}{10-4}(-1+0+0+1+2+2) = 4/6$
- $\bar{y}_{wk} = \frac{1}{10}[3(-1)+(0+0+1+2)+3 \times 2] = 3/5$

Slide 52 Stat 35, UCLA, Jon Dinger