BUSI 275: Business Statistics

- ***No food/drink in the computer lab, please!***
- ***Syllabus, etc. are on myCourses***
- ***Pre/co-req: ≥ C in MATH 101 or 120***

10 Jan 2012
Dr. Sean Ho

busi275.seanho.com

TRINITY WESTERN UNIVERSITY

Outline for today

- Welcome, devotional
  - Syllabus, text, myTWU, schedule
- Introduction to statistics for business:
  - Decision making, asking good questions
  - Variables: levels, IV/DV, cross-sectional
  - Stages/cycles in statistical analysis
- Term Project
- Exploring Data using Charts
  - For qualitative vars, for quantitative vars

What is statistics?

- Data-driven decision making
  - Evidence-based, not (only) “gut feeling”
- Answering vital ?s about business processes
  - “Which market segment is most price-conscious?”
  - “Which app model produces more revenue: free ad-funded or pay to play?”
- Asking more relevant questions
  - “How do we measure customer satisfaction?”
  - “What factors have the strongest influence on employee retention?”

Basic terms

- Population: group of interest
  - e.g., TWU students
- Sample: participants in our study
  - e.g., 50 passers-by at the cafeteria (time?)
  - Sampling: drawing a sample from the pop
  - Inferences: estimates (guess) on the pop
- Variable: measurable of interest
  - e.g., "monthly cellphone bill in $"
- Observation: values for a single participant
  - e.g., Jane's cell bill is $40/mo

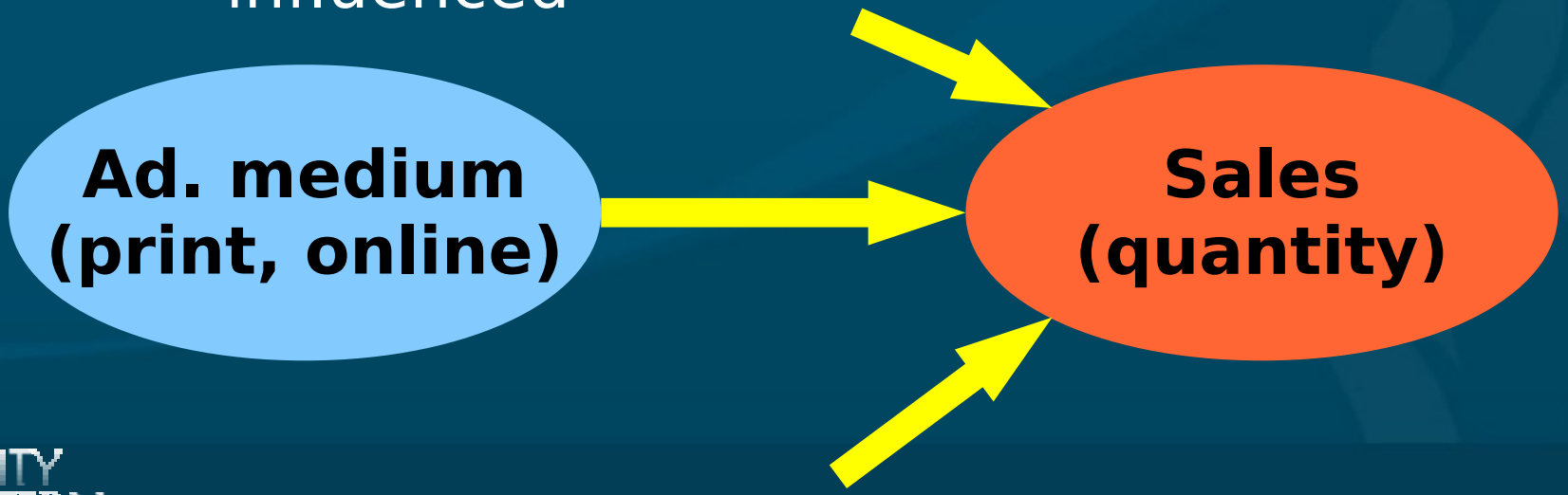
Levels of measurement

- Nominal (categorical):
  - Province, colour, store branch
    - Any yes/no (dichotomous) question
  - "Are you satisfied as a customer?"
- Ordinal (has an ordering, </> makes sense):
  - Letter grade, "satsifactory … unsatisfactory"
  - "very satisfied, somewhat satisfied, …"
- Interval (can do +/-/avg, but 0 is arbitrary):
  - ºC/F, Likert scale ("on a scale of 1-5")
- Ratio (can do mult/divide):
  - Salary, quantity of sales, height in cm

"Qualitative"

Direction of influence

- Statistical analysis often is about relationships amongst variables:
  - "Does advertising medium affect sales?"
- Often, one variable drives/influences another:
  - Predictor (independent variable, IV) drives
  - Outcome (dependent variable, DV) is influenced



Cross-sectional vs. time-series

- Cross-sectional data look at a snapshot in time:
  - e.g., 2011 revenue for various store branches
- Time-series data track the same variables on the same participants, at several points in time:
  - Annual revenue for branches, 2001-2011
- Time-series data need to worry about
  - Attrition (missing data)
  - Sampling in time (e.g., monthly vs. annual)
  - Uneven time (2010, 2009, and "<2009")
- Our class will mostly examine cross-sectional data

Cycles in statistical analysis

- Formulate research question (RQ)
- Gather data: sampling, metrics
- Prep data: input errors/typos, missing data, obvious outliers
- Explore variables: IV, DV, charts
- Model building: choose a model based on RQ
- Check assumptions of model
  - If not, either clean data or change model
  - May need to modify RQ!
- Run final model and interpret results

Research question: example

- RQ: are men taller than women?
  - Is this relationship real? How strong is it?
- What are the variables? IV/DV? Level of meas?
  - Levels of measurement: categorical, ordinal, scale (interval, ratio)
  - IV: gender (dichot), DV: height (scale)
- What type of test should we use?
  - Independent samples *t*-test
- Limitations/assumptions of this test?
  - We will learn about these: this is the point of this course!

Model-building process

- Operationally define a phenomenon: variables
- Measure it (collect data): how to do sampling?
- Build a model: verify data meet assumptions and input data into model
- Draw conclusions in the "real world" population
  - e.g., Child A has 2 apples, B has 6, and C has 1. How many apples is a child most likely to have?
  - Individual vs. group
- "Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted." (Albert Einstein)

e.g.: Retail duration of stay

- RQ: Does volume of music affect duration of customer stay in retail shops?
- Population, sample: how to gather data?
- Variables: how to measure?
  - Volume, in dBA
  - Duration of stay, in seconds
- Predictor (IV) / outcome (DV)?
  - Predictor: volume. Outcome: duration of stay
- Levels of measurement?
  - Volume (dBA): ratio. Duration (seconds): ratio
- Moving beyond: other predictors of duration?

Outline for today

- Welcome, devo, syllabus, myTWU, schedule
- Introduction to statistics for business
- Term Project:
  - Proposal
  - Dataset description
  - REB application
  - Presentation
  - Term Paper
- Exploring Data using Charts
  - For qualitative vars, for quantitative vars

Term Project

- A big part of this course is your term project:
  - Find suitable data:
    - Use existing (perhaps public) data, or
    - Collect your own (subject to REB approval)
  - Propose a statistical analysis of it
  - Get approval by our Research Ethics Board
  - Go through the “spiral” of statistical analysis
  - Write it up in an MLA-style manuscript
- Groups of up to 4 people
  - Email me when you have your group

Project Proposal (23Jan - 3Feb)

- Email me to setup your proposal meeting
  - All team members must be present
- ≥24hrs before the meeting, submit a written proposal for your project. Describe
  - Your population of interest
  - The key DV (outcome) and why it matters
  - Predictors that you think influence the DV
  - State specific research questions
  - Plan for how you will get your data
  - Plan for how you will divide the work amongst your team members

Dataset Description (due 7Feb)

- Use existing data, or gather your own
  - Getting data takes time! (and may need REB)
  - No simulated (made-up) data
- Minimum sample size: 80 (# observations)
- Remember the DV is the focus of your study
  - Gather relevant IVs to explain the DV
- Possible sources: your own data, faculty members, publicly available / government data (BLS, DataBC, etc.)
  - StatCan caveat: getting the original microdata is tricky, sometimes costs $

REB application (due 14Feb)

- Approval by TWU Research Ethics Board is required before any new analysis may be done!
- You are not allowed to start your analysis until you get REB approval (expect 3-6 weeks)
  - Can't even recruit your study subjects!
- Use either the "Request for Ethical Review" form or the "Analysis of Existing Data" form
- For existing non-public data, you need written permission from the original owner of the data
- Some data, e.g., public datasets, may be REB exempt

Presentations (10Apr)

- 15-min in-class presentation
- Target audience may not care about stats
  - e.g., your company's CEO or board
  - Give practical recommendations or lessons-learned
- Motivate why we should care about your topic
- Have some preliminary results to show
- Every team member must participate
- Also complete feedback forms for other teams' presentations

Term Paper (due 16Apr)

- Aim at non-statistician (CEO, etc.)
  - But back up your conclusions with statistical results, using APA style or similar
  - Include enough details to reproduce study
- Proper, professional English
  - Format in MLA, APA, or similar style
- Related work / background research
  - Cite references
- Include relevant figures / tables
  - Can include more in appendix or separate Excel

Outline for today

- Welcome, devo, syllabus, myTWU, schedule
- Introduction to statistics for business
- Term Project and milestones
- Exploring Data using Charts (ch2)
  - Frequency distributions
    - Working with Excel array formulas
  - Crosstabs / pivot tables
  - Histograms and the ogive
  - Scatterplots
  - Line charts

- ***Please download from lecture dir: 01-SportsShoes.xls***

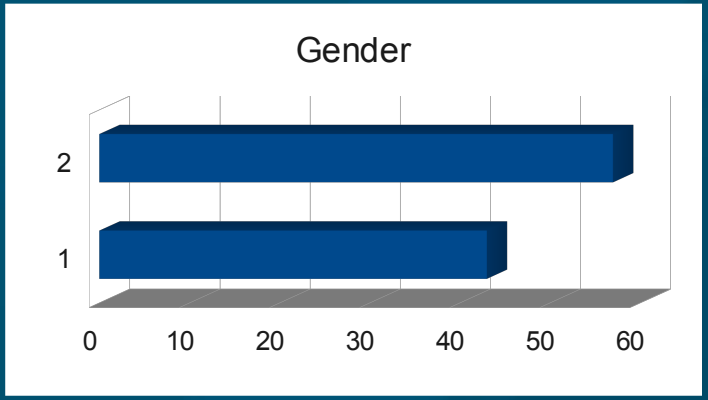
Frequency distributions

- How frequently each value of a variable appears in the dataset (either pop or sample)
- Data usually come as 1 row = 1 participant:
  - Sample size = #rows

1	Homeroom #	First Name	Last Name	Payment	T-Shirt Color	T-Shirt Size
3	105	Esther	Yaron	7-Oct	Dark Red	Small
4	105	Melissa	White	7-Oct	Heather Grey	Small
5	220-A	Christopher	Peyton-Gomez	Pending	White	Small
6	220-A	Brigid	Ellison	Pending	Dark Red	Small
7	220-B	Windy	Shaw	7-Oct	Heather Grey	Small
8	220-B	Malik	Reynolds	7-Oct	Heather Grey	Small
9	220-B	Michael	Lazar	14-Oct	White	Small
10	105	Christiana	Chen	5-Oct	Dark Red	Medium
11	105	Sidney	Kelly	11-Oct	Dark Red	Medium
12	105	Nathan	Albee	13-Oct	Heather Grey	Medium
13	110	Matt	Benson	11-Oct	White	Medium
14	110	Gabriel	Del Toro	13-Oct	White	Medium
15	135	Chantal	Weller	15-Oct	White	Medium

- Compute by tallying up how many occurrences of each value exist in the data:
  - e.g., for "T-Shirt Size" *(level of meas?)*:
    Small: 10; Medium: 20; Large: 15

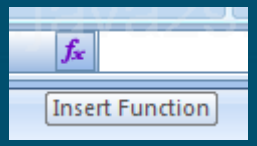
Excel: freq. dist. & bar chart



- Dataset: 01-SportsShoes.xls
  - Add new sheet: "Charts"
- Frequency distribution:
  - Enter poss. values (Gender coded as 1, 2)
  - Highlight range of cells for output
  - Input the FREQUENCY() Excel function
  - Array formula: use Ctrl-Shift-Enter
  - To get relative freqs (%), divide by total
    - Absolute cell refs: 'A2'
- Bar chart: Insert > Bar > 2D > Select Data:
  - Data: freqs; Cat. Axis Labels: values

Excel array formulas

- Regular formulas (functions) take cells or cell ranges as input and produce a single output
  - Array formulas output to a range of cells
- Highlight the range where output will go
- Enter the formula:



  - =FREQUENCY()
  - Data: highlight Data!M2:M101
  - Bins (values): highlight cells with “1”,”2”
- Don't hit OK yet! Use Ctrl-Shift-Enter instead to indicate it is an array formula

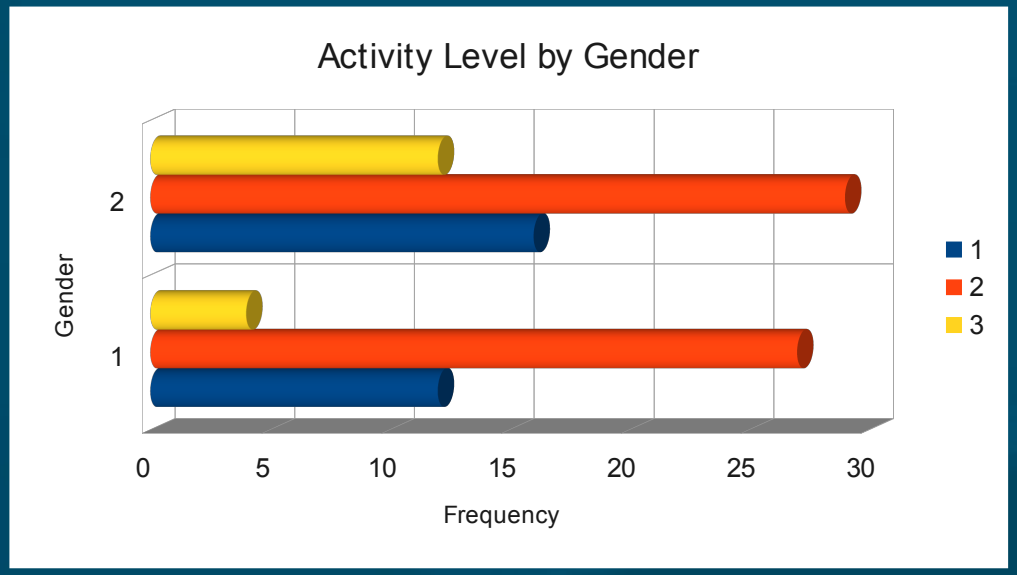
Multiple vars: crosstabs

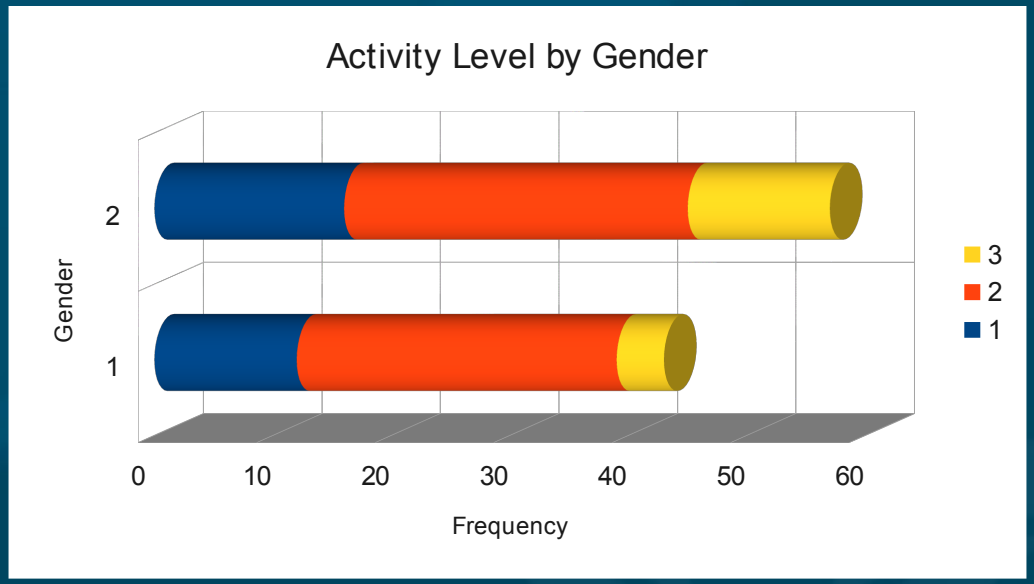
- Consider all combinations of values:
  - e.g., Gender: 1 or 2; Activity: 1, 2, 3
    so there are 6 combos of (Gender, Act)
- Cross-tabulations (Pivot Tables, Joint freq. dist):
  - Insert > Pivot Table
  - Select Range: L1:M101
  - Row Labels: Gender
  - Col Labels: Activity
  - Values: either
  - Summarize By: Count

Count - Activity	Activity			
Gender	1	2	3	**Total Result**
1	12	27	4	**43**
2	16	29	12	**57**
Total Result	**28**	**56**	**16**	**100**

Multiple vars: clustered bars

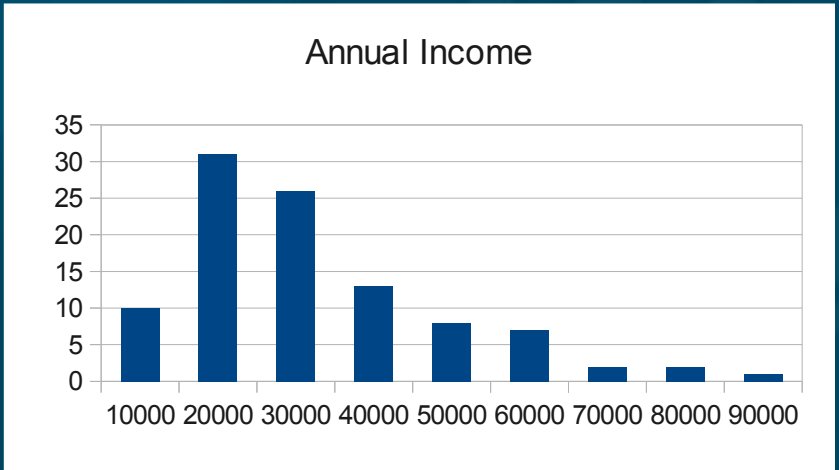
- If one of the nominal variables only has a few possible values (categories), then
- We can use clustered or stacked bar charts:





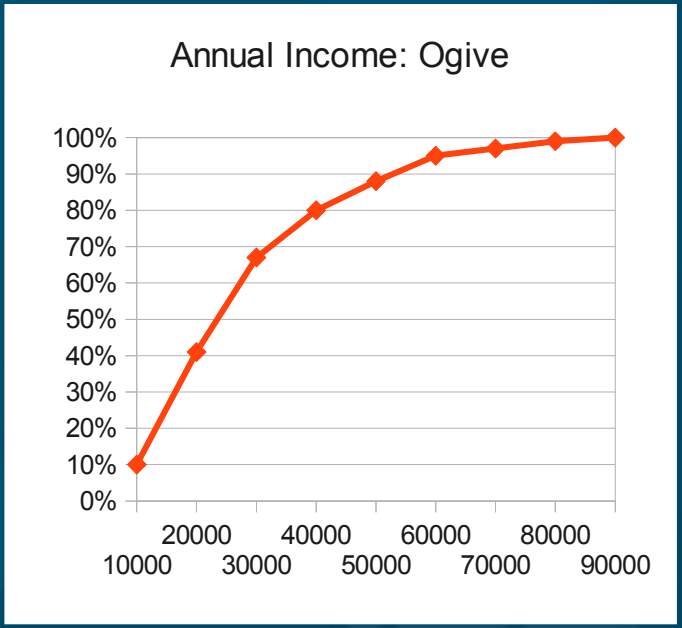
Quantitative vars: histograms

- For quantitative vars (scale, ratio), must group data into classes
  - e.g., length: 0-10cm, 10-20cm, 20-30cm... (class width is 10cm)
  - Specify class boundaries: 10, 20, 30, ...
- How many classes? for sample size of n, use k classes, where $2^k \geq n$
- Can use FREQUENCY() w/ column chart, or
- Data > Data Analysis > Histogram



Cumulative distrib.: ogive

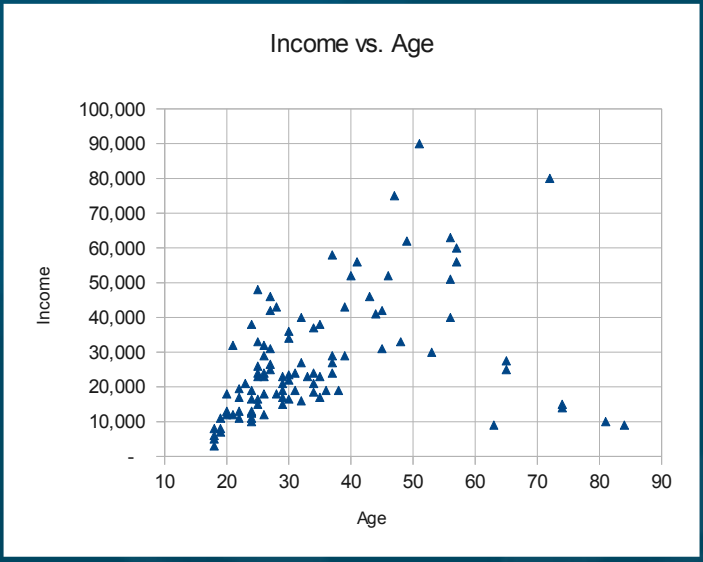
- The ogive is a curve showing the cumulative distribution on a variable:
  - Frequency of values equal to or less than a given value
- Compute cumul. freqs.
- Insert > Line w/Markers



- Pareto chart is an ogive on a nominal var, with bins sorted by decreasing frequency
  - Sort > Sort by: freq > Order: Large to small

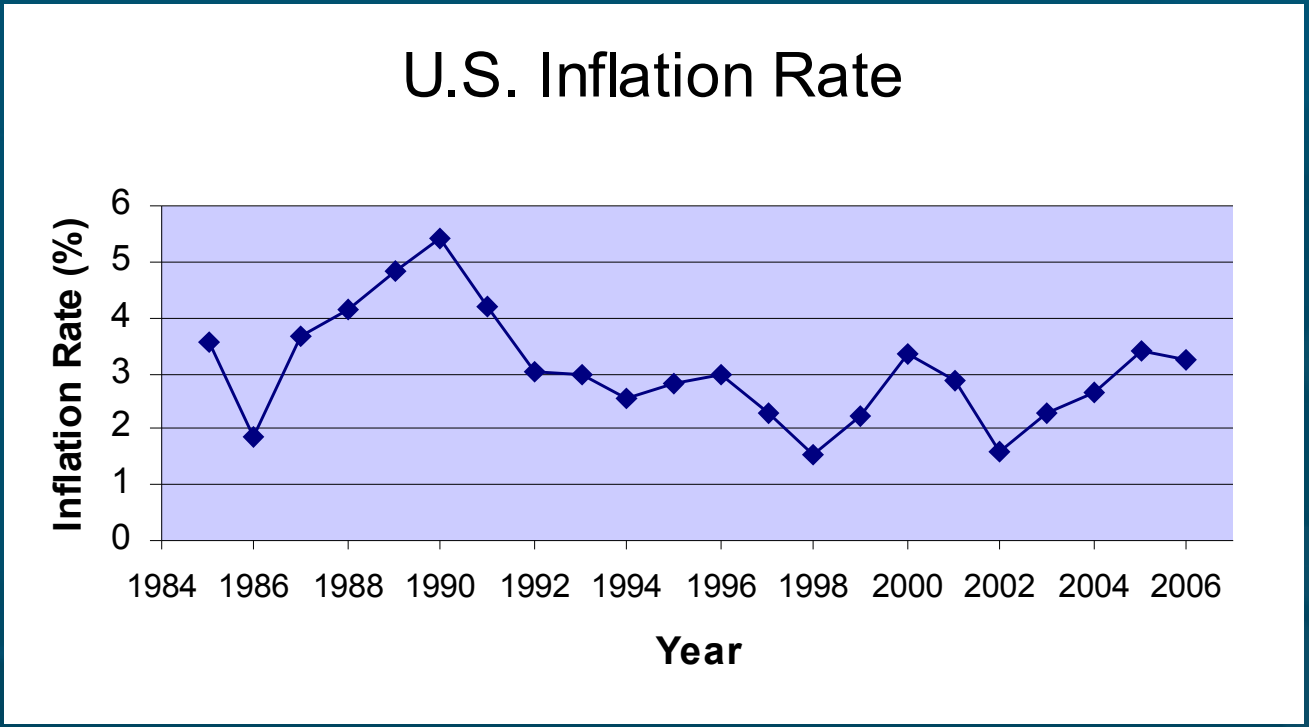
2 quant. vars: scatterplot

- Each participant in the dataset is plotted as a point on a 2D graph
  - (x,y) coordinates are that participant's observed values on the two variables
- Insert > XY Scatter
- If more than 2 vars, then either
  - 3D scatter (hard to see), or
  - Match up all pairs: matrix scatter



Time series: line graph

- Think of time as another variable
  - Horizontal axis is time
- Insert > Line > Line



TODO

- HW1 (ch1-2): due next week Thu 19Jan
  - Text document: well-formatted, complete English sentences
  - Excel file with your work, also well-formatted
  - HWs are to be individual work
- Get to know your classmates and form teams
  - Email me when you know your team
- Discuss topics/DVs for your project
  - Find existing data, or gather your own?
- Schedule proposal meeting during 23Jan - 3Feb