Statistical Methods for BCS

9.07

9/8/2004

BCS 9.07

- MWF 10-11am
- Instructor: Ruth Rosenholtz
  - Office hours W 3-4 pm
- Textbook:
  - Basic Statistics for the Behavioral Sciences, Gary W. Heiman, 4th edition
  - There will likely also be outside readings, available on the MIT server

- PowerPoint lectures will generally appear by 9 pm the day before lecture.
  - This is contingent on class attendance!
- TA/instructor contact info, tentative schedule, homework, and extra handouts will also appear on the MIT server.
  - Your first homework is there now. Due Friday next week.
  - You can turn in homework on the MIT server, but are not required to do so.

Grading

Class participation	5%
Homework*	40%
Mid-term	20%
Final	35%

*We’ll be using MATLAB for much of the homework. It is available on the MIT server cluster machines.

Please turn in a copy of your MATLAB code with each assignment. This helps us to assign partial credit if you make an error.

Academic honesty policy

- You may work with others on your homework
  - I.E. you may discuss your homework with other students
  - Write, at the top of your HW, “I worked with…”
  - But, each student must solve each problem themselves (including writing their own MATLAB code), and write up the solutions themselves
  - It is *never* acceptable to copy from someone else’s solution
  - If it seems you have copied from someone else’s solution, you will get 0 points for that problem, and we will go over the rest of your HW with a fine-tooth comb
  - Since it is often difficult to tell who copied from whom, *don’t let anyone else copy off of your homework!*

Academic honesty policy

- You are expected *not* to make use of solutions or assignments from previous years
- Obviously, don’t copy off anyone else’s exam, either

Policy on late homework

- *Short* extensions may be granted due if you have a reasonable excuse (MIT sporting event, wedding, job interview, etc.), provided you notify us *on or before the day the homework is assigned*.
  - However, note that you will receive homework assignments nearly 2 weeks before they are due, so just being out of town for a day or two may not be a sufficient excuse.
- For more unforeseen difficulties, an extension may also be granted, provided you get a letter from your doctor, a Dean, Counseling & Support Services, or equivalent.
  - However, note that we may be limited in our options if we learn of your difficulty too late. Please let us know informally that there may be a problem as soon as you know of it (not 2 weeks after the assignment was due, and not at the end of the semester!).

Statistics

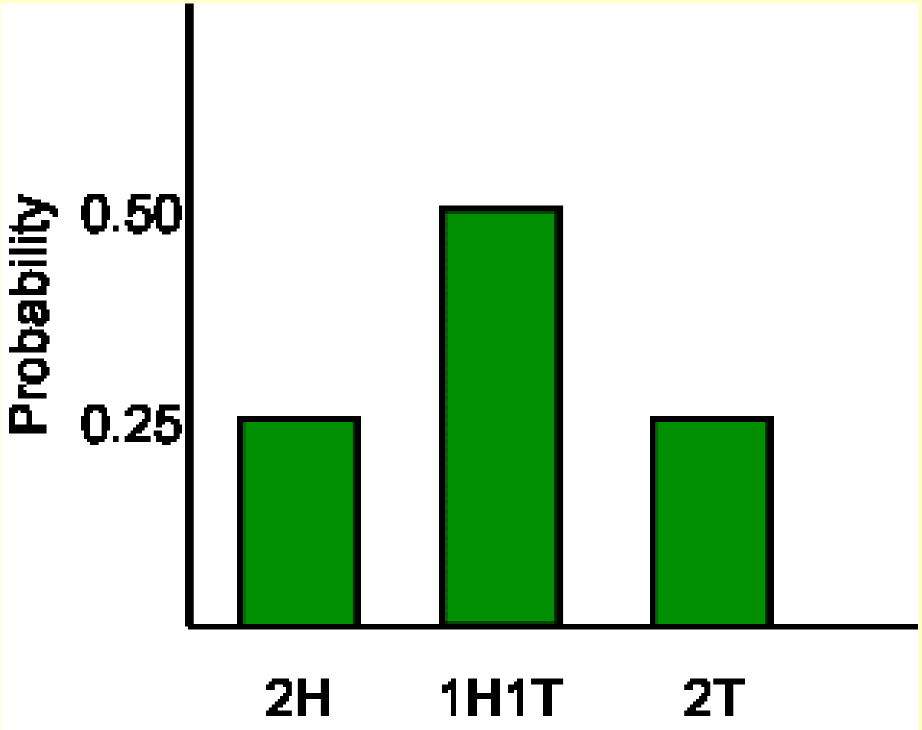
Flipping two fair coins

- What is the probability of getting two heads?
  - 0.5*0.5=0.25
- What is the probability of getting two tails?
  - 0.5*0.5=0.25
- What is the probability of getting one head and one tail?
  - 0.5 (head) * 0.5 (tail) + 0.5 (tail) * 0.5 (head) = 0.5

If you don’t remember how to do this, don’t worry, we’ll review probability next week.

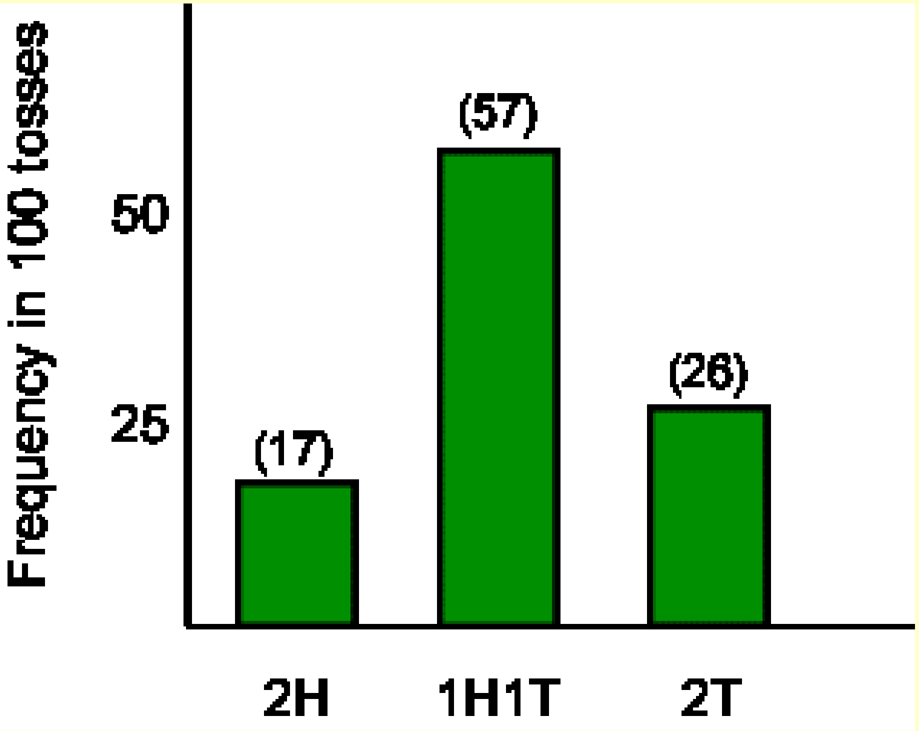
Probability density function (PDF)

- Represents the true (in this case, theoretical) probability of occurrence of the set of possible events.



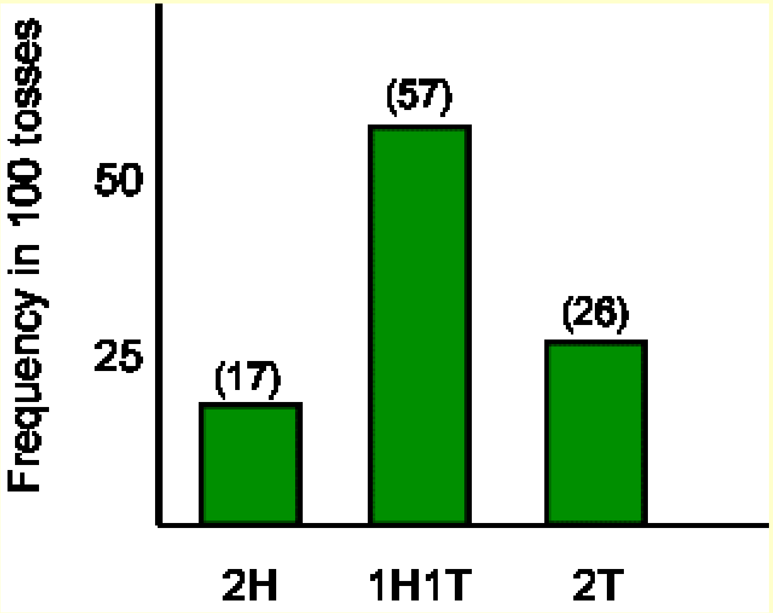
Frequency histogram

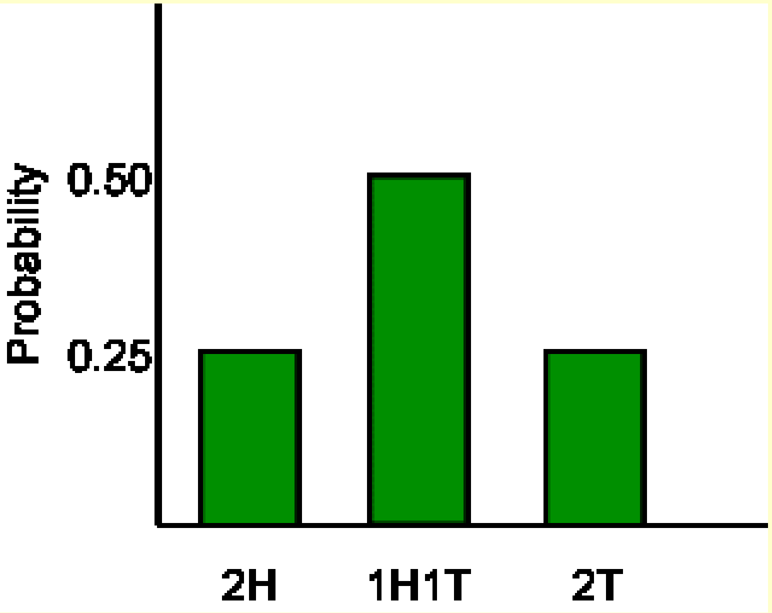
- Represents the actual frequency of occurrence of events in a *sample*.
- Here, I flipped a pair of coins 100 times.



Are my 2 coins fair coins?

- Frequency histogram
- PDF



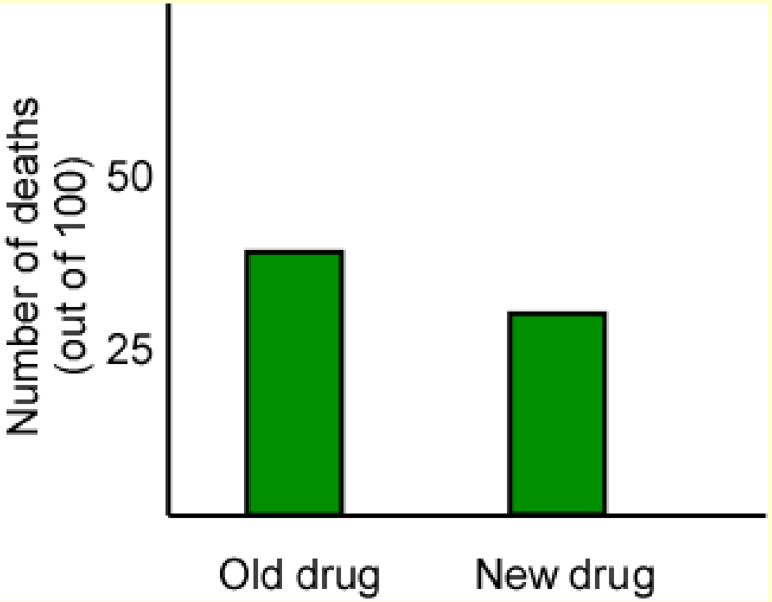


Are my 2 coins fair coins?

- The frequency histogram doesn’t quite match the pdf.
- It’s difficult to tell from the data whether this is due to a *systematic factor* (unfair coins) or *chance* (or both). This is where statistics comes in.
- In this case, the coins are fair (the coin flips were generated in MATLAB).
- We only expect the distribution of coin flips to look like the pdf in the long run. Not in a particular sample of 100 flips.
- An outcome can differ from what is expected just by chance.

Does a new drug cure cancer better than the old drug?

- The data:

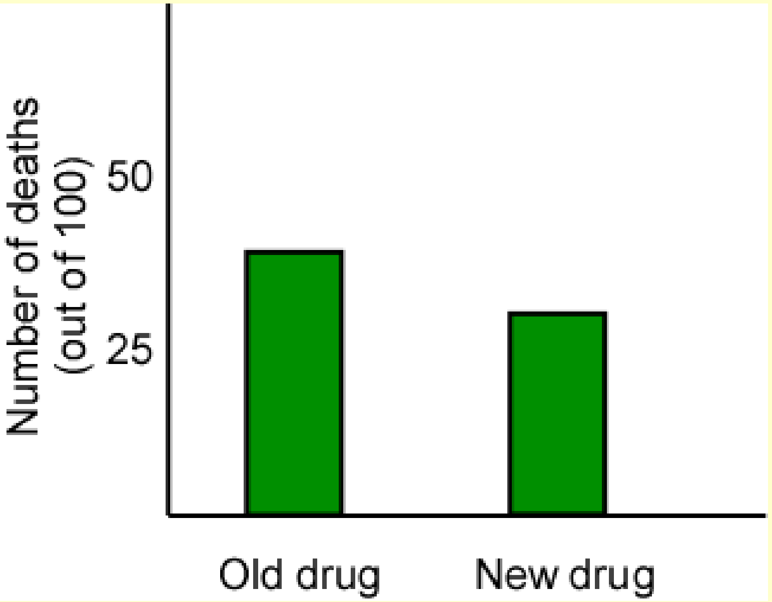


Does a new drug cure cancer better than the old drug?

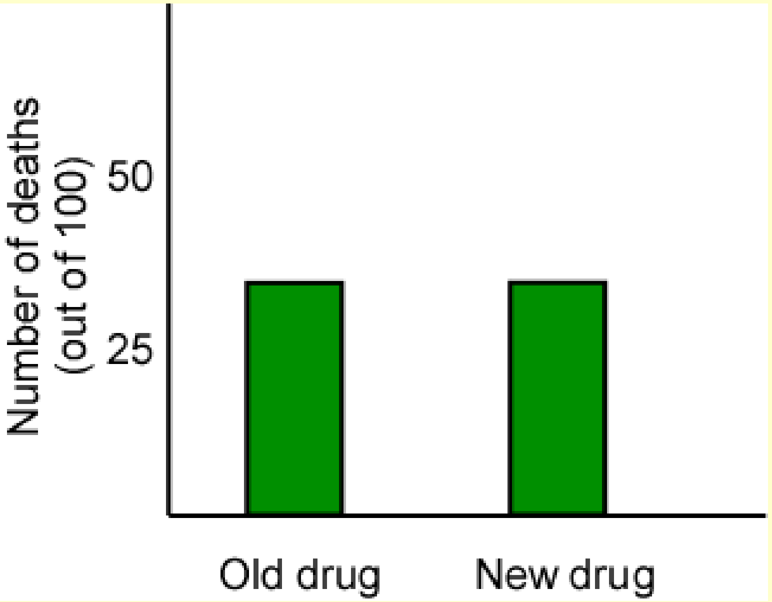
- There's an empirical difference between the old drug and the new drug, but is it due to a systematic factor (e.g. the new drug works better) or due to chance?
- A related question: if we gave this drug to 100 more people, would we expect to continue to see improvement over the old drug? Do we expect this effect to *generalize*?

Alt: Is the difference between data & theory due to systematic factors + chance, or to chance alone?

- Data:



- "Theory" = no difference between the drugs



Chance vs. systematic factors

- A *systematic* factor is an influence that contributes a predictable advantage to a subgroup of our observations.
  - E.G. a longevity gain to elderly people who remain active.
  - E.G. a health benefit to people who take a new drug.
- A *chance* factor is an influence that contributes haphazardly (randomly) to each observation, and is unpredictable.
  - E.G. measurement error

Observed effects can be due to:

A. Systematic effects alone (no chance variation).

- We’re interested in systematic effects, but this almost never happens!

B. Chance effects alone (all chance variation).

- Often occurs. Often boring because it suggests the effects we’re seeing are just random.

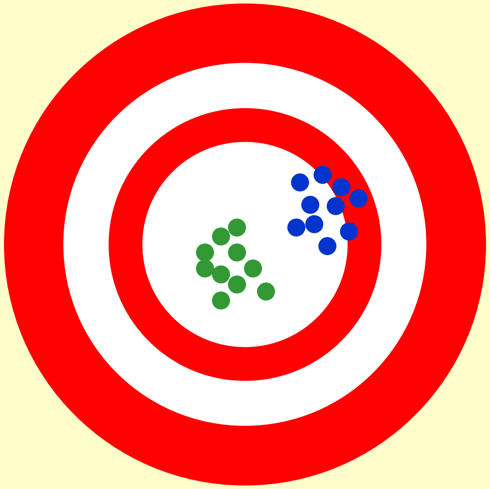
C. Systematic effects plus chance.

- Often occurs. Interesting because there’s at least some systematic factor.

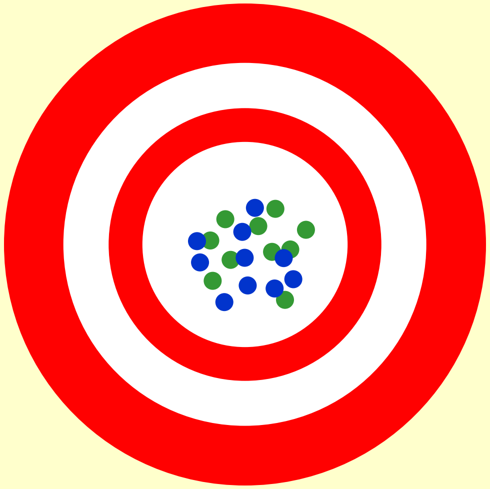
An important part of statistics is determining whether we’ve got case B or C.

Systematic + chance vs. chance alone

- Likely systematic + chance variation:



- Likely due to chance alone:



No chance variation

No chance variation is like when Robin Hood shoots his second arrow in exactly the same place as his first, so the second one splits the first arrow down the middle!

On a scale from 1 to 10, rate your experience at MIT so far:

7, 7, 7, 7, 7, 7, 7, 7, 7, 7…

We have a natural tendency to over-estimate the influence of systematic factors

- How well a baseball player does in a given at-bat depends on both chance and the skill of the batter. How much of each?
- I.E. there's some amount of variation from at-bat to at-bat, regardless of whether it's a new batter or the same batter trying again. What percent of total variation is accounted for by changing batters?
  - True: <0.5% of the variation is due to differences in skill of different batters.
  - But baseball fans estimated about 25% was due to differences in skill. (Abelson, 1985)

We have a natural tendency to over-estimate the influence of systematic factors

- The lottery is entirely a game of chance (no skill), yet subjects often act as if they have some control over the outcome. (Langer, 1975).
- We tend to feel that a person who is grumpy the first time we meet them is fundamentally a grumpy person. (The "fundamental attribution error," Ross, 1977.)

- As researchers, we need a principled way of analyzing data, to protect us from inventing elaborate explanations for effects in data that could have occurred predominantly due to chance.
- This is what statistics is for.

Some Definitions

- Sample
  - A group of individuals, or the data (scores) from those individuals, chosen to represent a larger population
    - This is the data we gathered in our experiment.
- Population
  - A group of all individuals who share some common feature or features, or the data (“scores”) one would obtain from that group of individuals
- Generalization
  - The process by which information about the population is inferred from the sample

Why statistics?

- We may see a difference between two conditions in a *sample*. Is it wise to *generalize* to the full *population?* Will the effect be reproducible if we run the experiment again on a different sample from the same population?
- When is error due to chance (measurement error, natural variability) instead of due to something systematic? We tend to be interested in systematic differences. If the effect is systematic, the experiment should be reproducible.

Example

- We test a new drug and a placebo on a *sample* of 100 U.S. females with depression. We find the subjects trying the new drug are 50% more likely to report a reduction in symptoms.
- Is it likely that this result generalizes to the *population* of U.S. females with depression?
- A combination of statistics, good experimental design, and additional experiments can help us answer this question.

Standalone statistics

- "The average life expectancy of famous orchestral conductors is 73.4 years." (Atlas, 1978)
  - Is 73.4 years unusual? Is it high or low?
- "Adults who watched television 3-4 hours a day had nearly double the prevalence of high cholesterol as those who watched less than one hour a day." (Tucker & Bagwell, 1992)
  - Does "nearly double" mean it's a bad idea to watch TV?

The importance of comparison

- With what do we compare famous conductors?
  - Conductors that are not famous?
  - The general population?
- The study compared with the life expectancy of males in the U.S. = 68.5 years. This is about a 5 year life extension, which seems a big enough difference to be important.

What might have *caused* this difference?

- Conductors do live longer.
  - Arm movements are good for you.
  - Health benefits from controlling others.
  - Common genetic basis to musical talent and longevity.
- The comparison standard is flawed.
  - Does the general U.S. population include people with shorter life spans who are also ineligible to become conductors? (E.G. people who are chronically ill?)
  - If so, the shorter life span of that population is likely a result of an inappropriate comparison standard rather than an effect of being a conductor.
- The difference is due to chance
  - By chance, the given sample of conductors lived particularly long.

Testing candidate explanations

- Conductors do live longer.
  - Select a sample of people. Randomly instruct half of them to move their arms like a conductor for some set amount of time per day. Do they live longer?
  - This is called a *randomized controlled* experiment. We'll talk about this more when we talk about experimental design.
- The comparison standard is flawed.
  - E.G. note life expectancy data for general public includes infant mortality – none of those people had a chance to become conductors. Carroll (1979) suggested comparing with life expectancy for U.S. males that have reached age 32. LE = 72.0 years.
- The difference is due to chance
  - Use statistics!
  - Determine how likely it is that the longevity difference is due to chance, as opposed to some *systematic* factor related to conducting.

Statistics as the Science of Understanding Data

- General goals of social science research:
  - Describe the data
  - Use the data to predict behavior
  - Determine the causes of behavior
  - Understand or explain behavior

Descriptive vs. Inferential Statistics

- Descriptive Statistics:
  - Describe the characteristics of a *sample* (usually drawn from a larger *population*)
- Inferential Statistics:
  - Infer characteristics of the larger population by *generalizing* from a representative sample of the population.

Uncertainty & incomplete information

- Throughout our lives, we make decisions based upon uncertain and incomplete information:
  - Should I order the veggie pizza? It's been really good 8 out of 10 times I've had it, but 2 out of 10 times it's been disappointing.
  - Should I try a new drug therapy?

Statistics & uncertainty

- Statistics does *not* allow us to *eliminate* that uncertainty.
  - ~~“I’m certain that the new drug therapy works.”~~
- But, it allows us to *quantify* it, which can help us make decisions.
  - “I’m 95% confident that the new drug therapy leads to 25-38% improvement over the old therapy.”

Descriptive statistics

- Inspecting the data visually, or graphing it.
- Describing the central tendency of the data.
- Describing the spread of the data.
- Describing the relationship between two sets of data.

Variables

- Anything that changes, in a study.
- E.G. a population survey asks: How old are you? What is your family's income? Do you have a job?
- Anything that varies from person to person on this survey is a variable. E.G:
  - Age, income, employment status.

Variables

- In an experiment, a variable could be something you measure, or a condition you change from trial to trial.
- E.G. you vary the amount of alcohol that you give to subjects, and measure their reaction time in a video game.
  - The amount of alcohol is an explanatory, or *independent*, variable.
  - The reaction time is a response, or *dependent*, variable.
- Explanatory (Independent) Variables
  - Variables whose values do not depend on the values of other variables in the data set
- Response (Dependent) Variables
  - Variables whose values are dependent on the values of other variables in the data set

The right sort of descriptive statistics depends upon the kind of data

- Quantitative variables
  - Age
  - Family size
  - Income
  - GPA
- Qualitative or categorical variables
  - Marital status
  - Employed or no?

More on quantitative variables

- Discrete: can vary by fixed amounts
  - Family size
- Continuous: can vary by arbitrary amounts
  - The age of two people can differ by years, a month, an hour,…

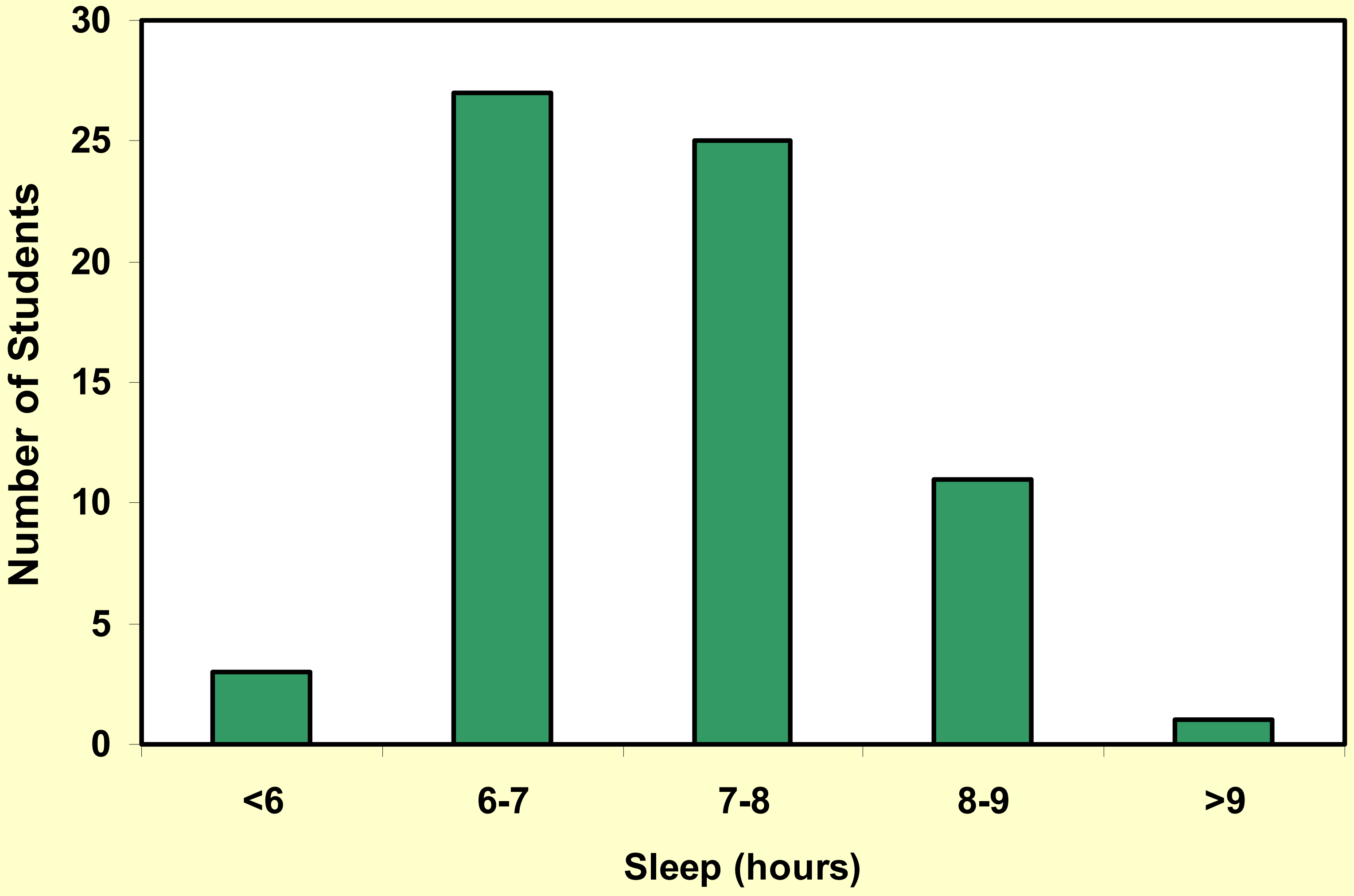
Types of Measurement Scales

- Four basic types
  - Nominal
    - Every element has a unique value -- no sense of quantity or order
    - E.G. telephone numbers, blood type
  - Ordinal
    - Like nominal, but with the addition of a sense of order
    - E.G. military rank, year of college, letter grades
    - Differences between adjacent values are not meaningful

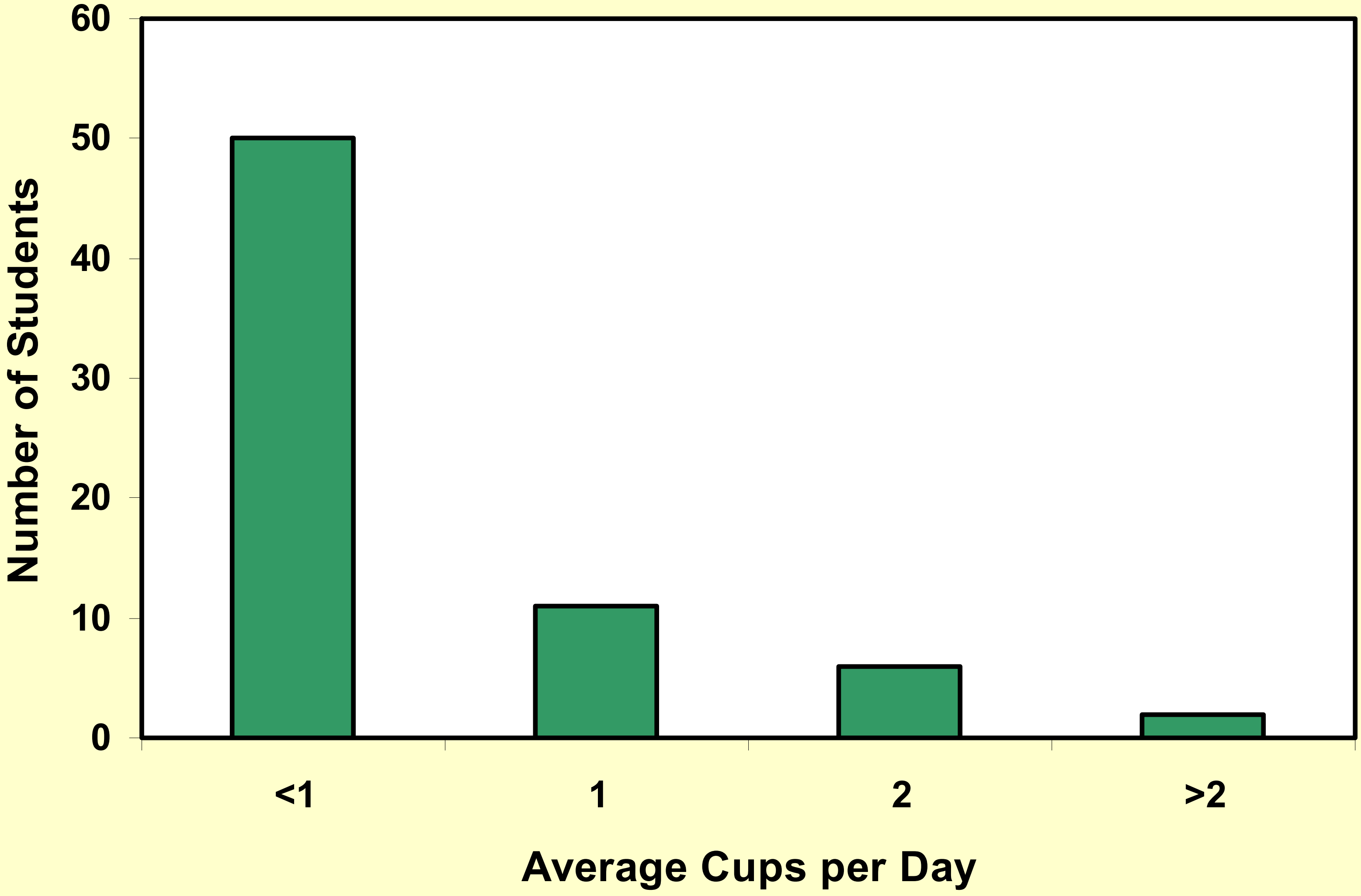
Types of Measurement Scales

- Four basic types
  - Interval
    - Like ordinal, with the addition of a unit of distance.
    - Can include negative numbers (not meaningful 0 point). Inappropriate to make "ratio statements", e.g. $4^{o}C \neq 2*2^{o}C$
    - E.G. temperature in Fahrenheit
  - Ratio
    - As above, plus the addition of a meaningful zero point
    - E.G. weight, height, distance traveled

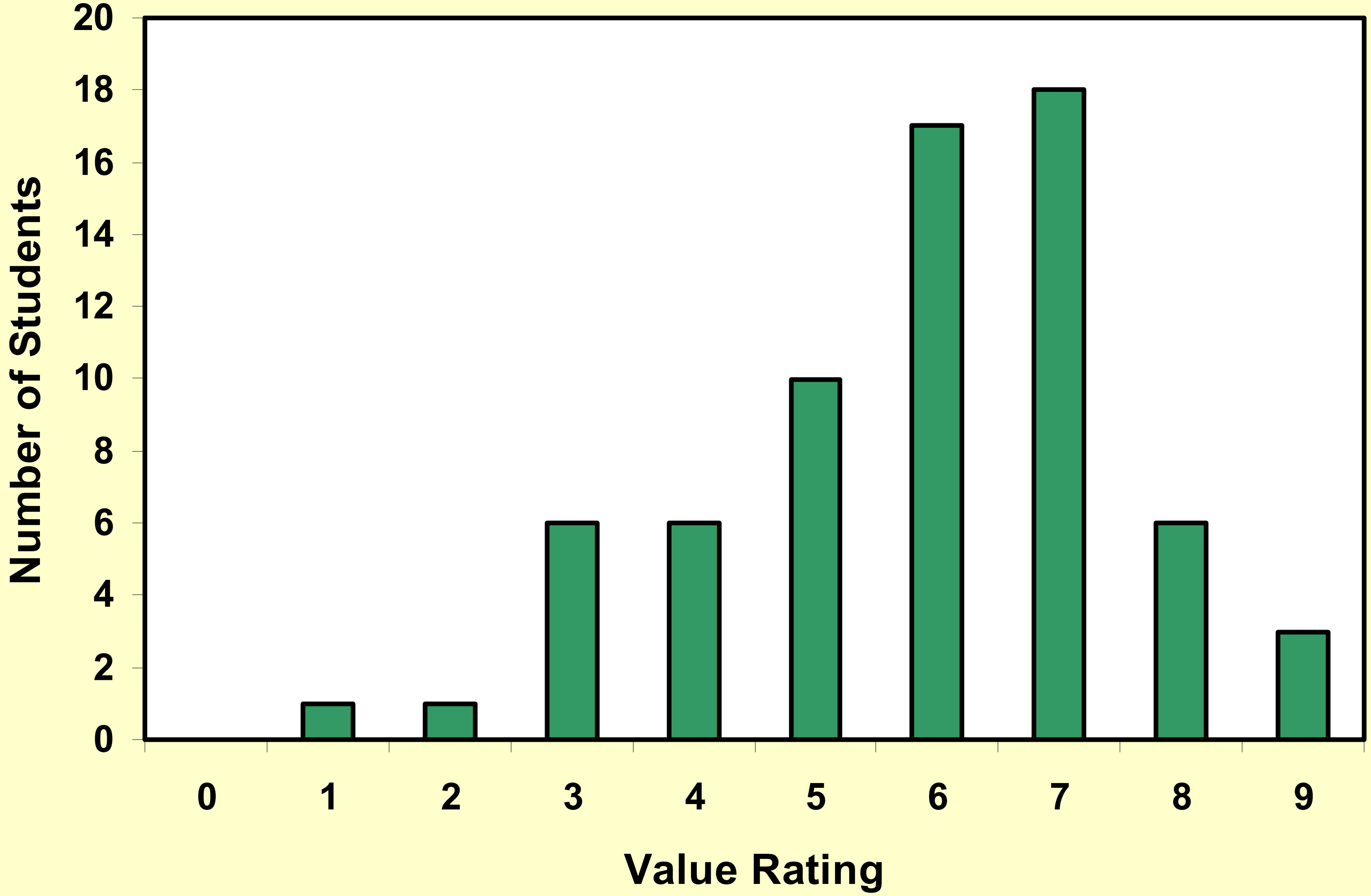
Hours of Sleep



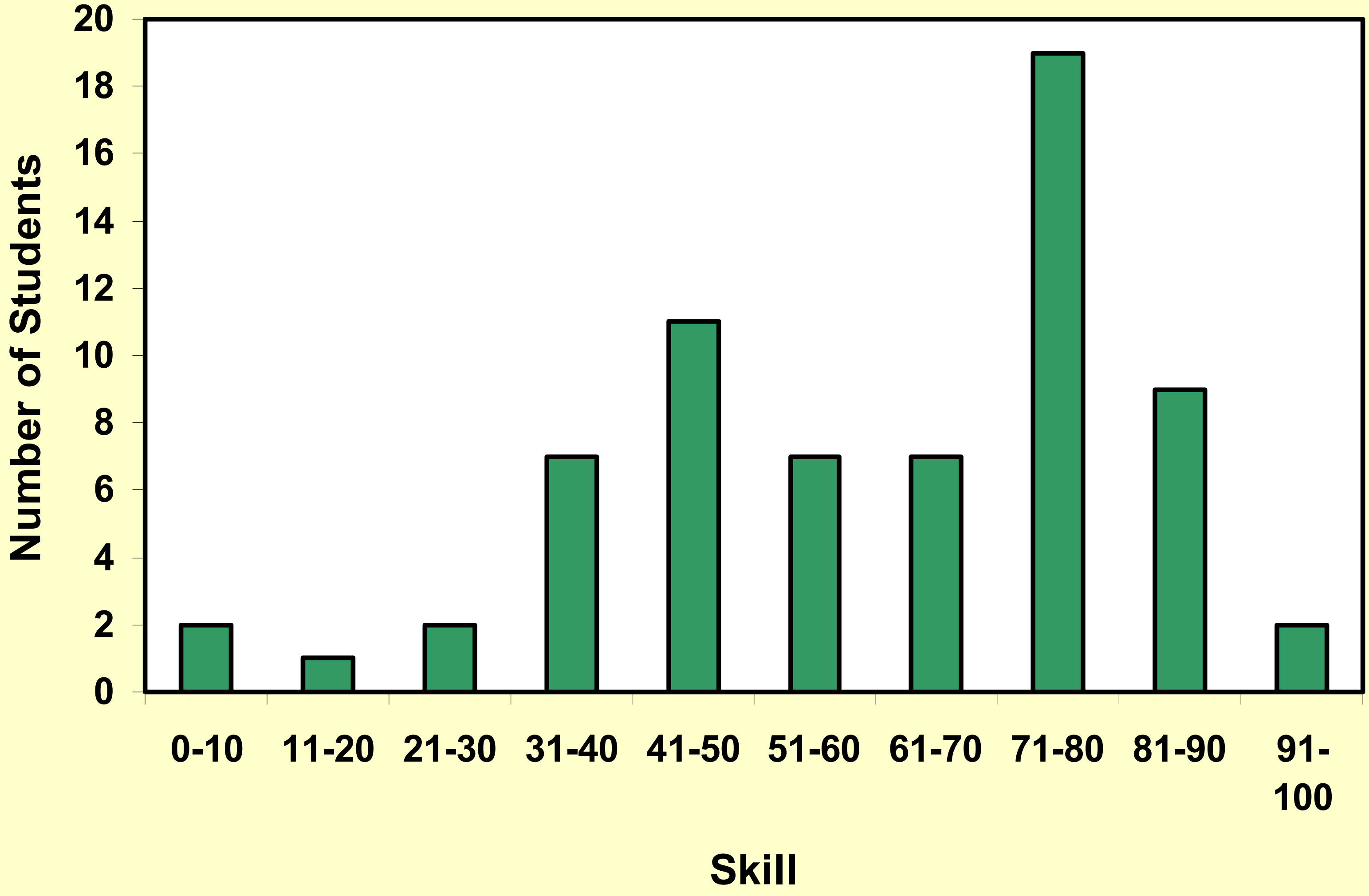
Cups of Coffee



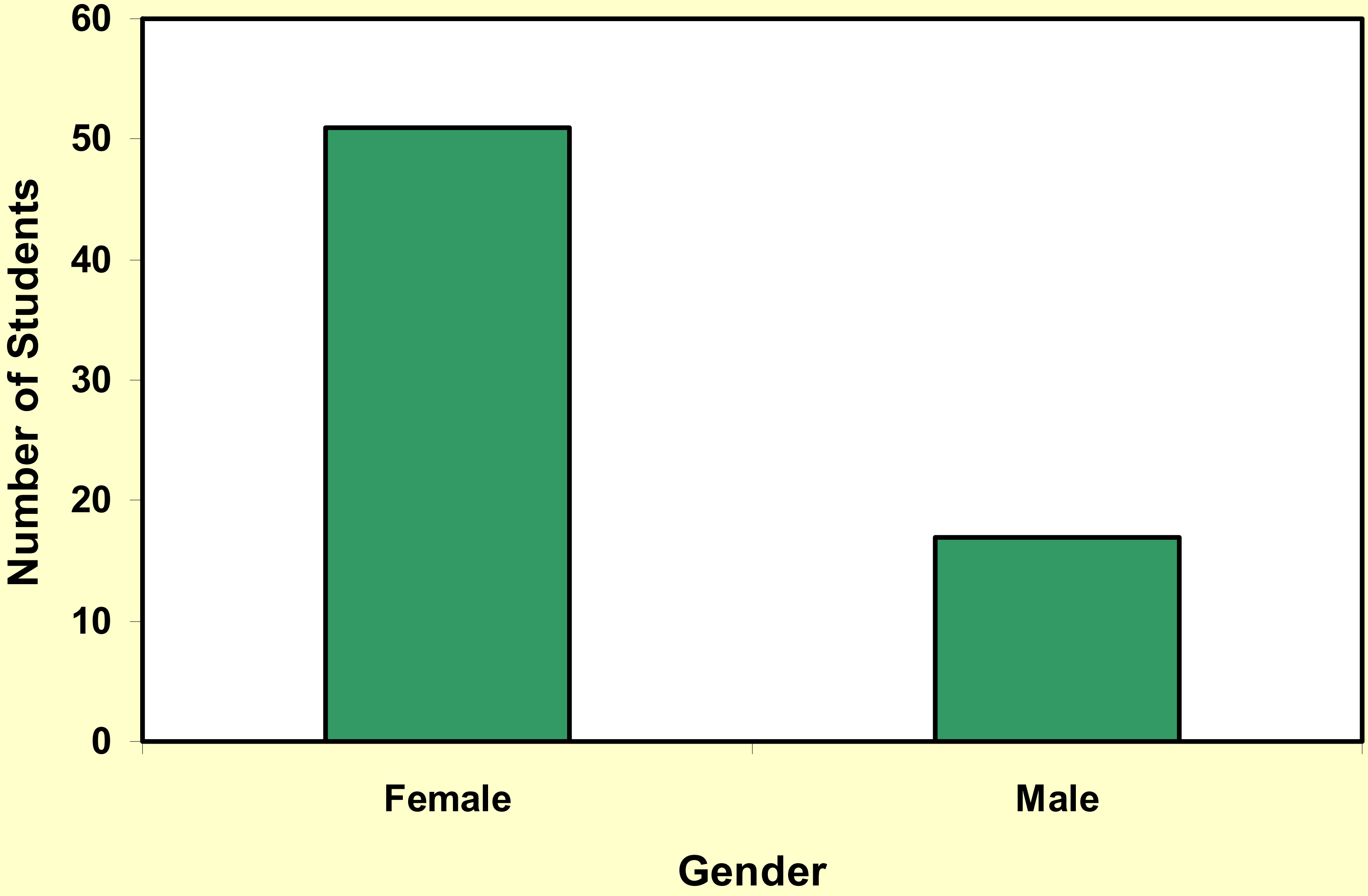
Importance of Statistics



Driving Skill (Self Rated)



Gender



Gender

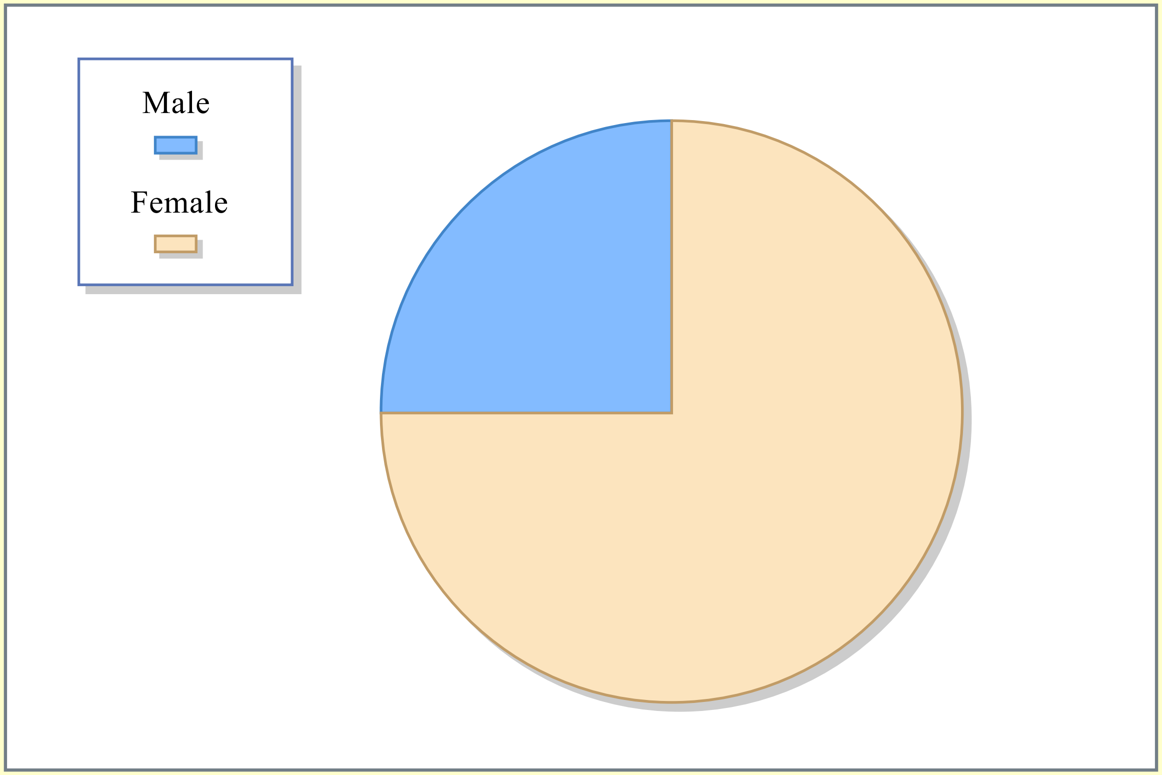
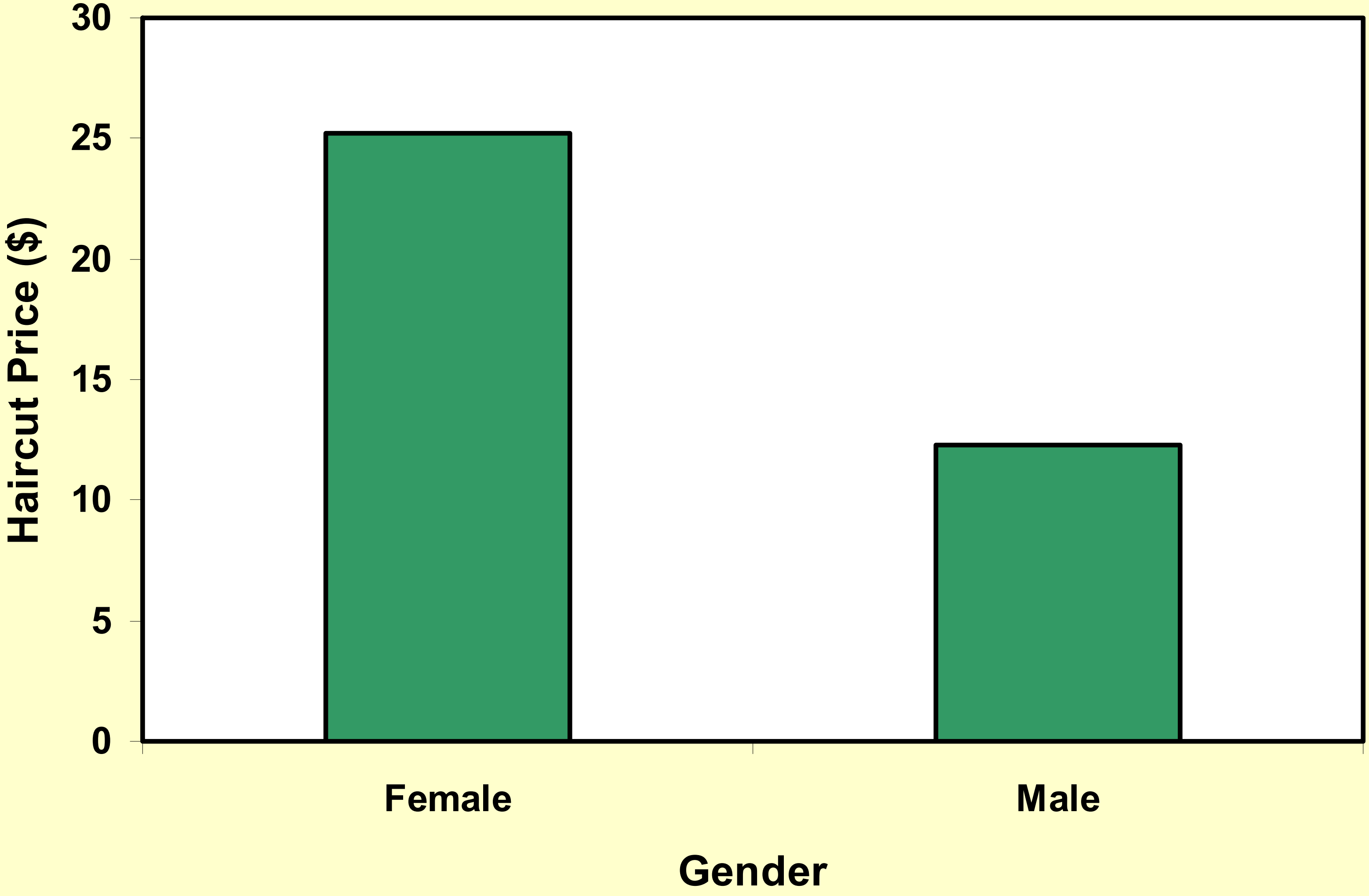


Figure by MIT OCW.

Haircut Costs



A bad graph...

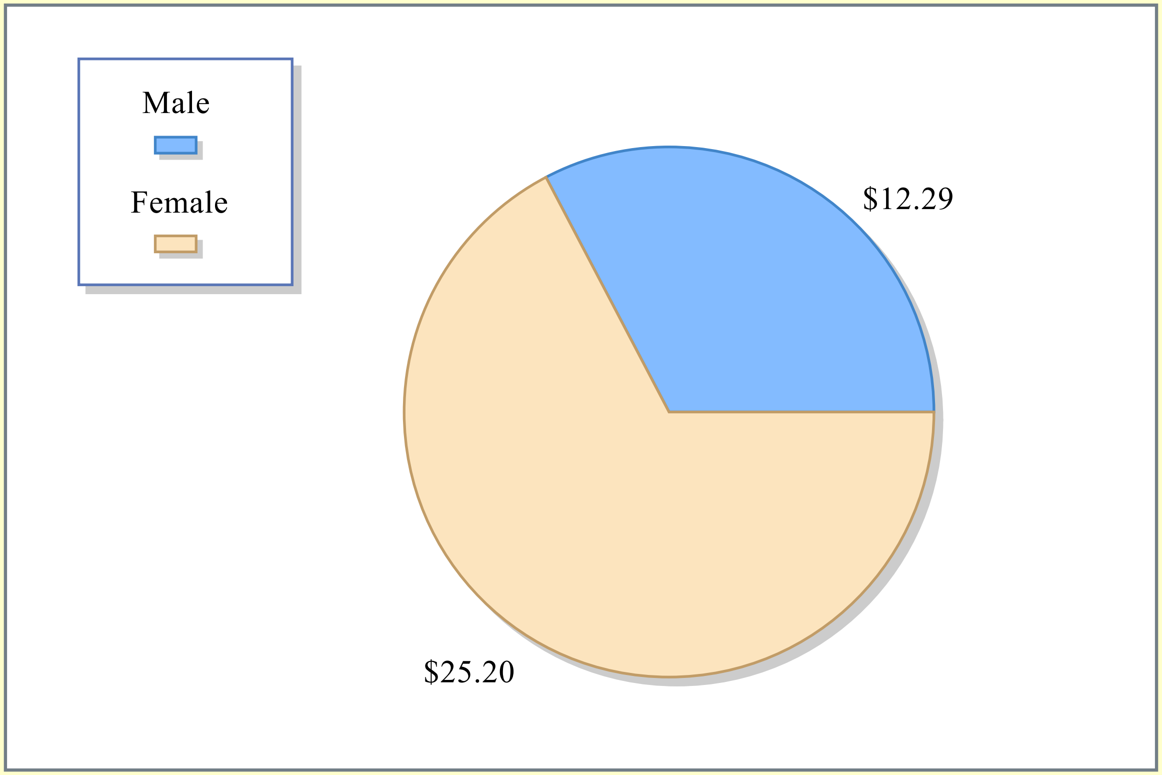


Figure by MIT OCW.

Political Views

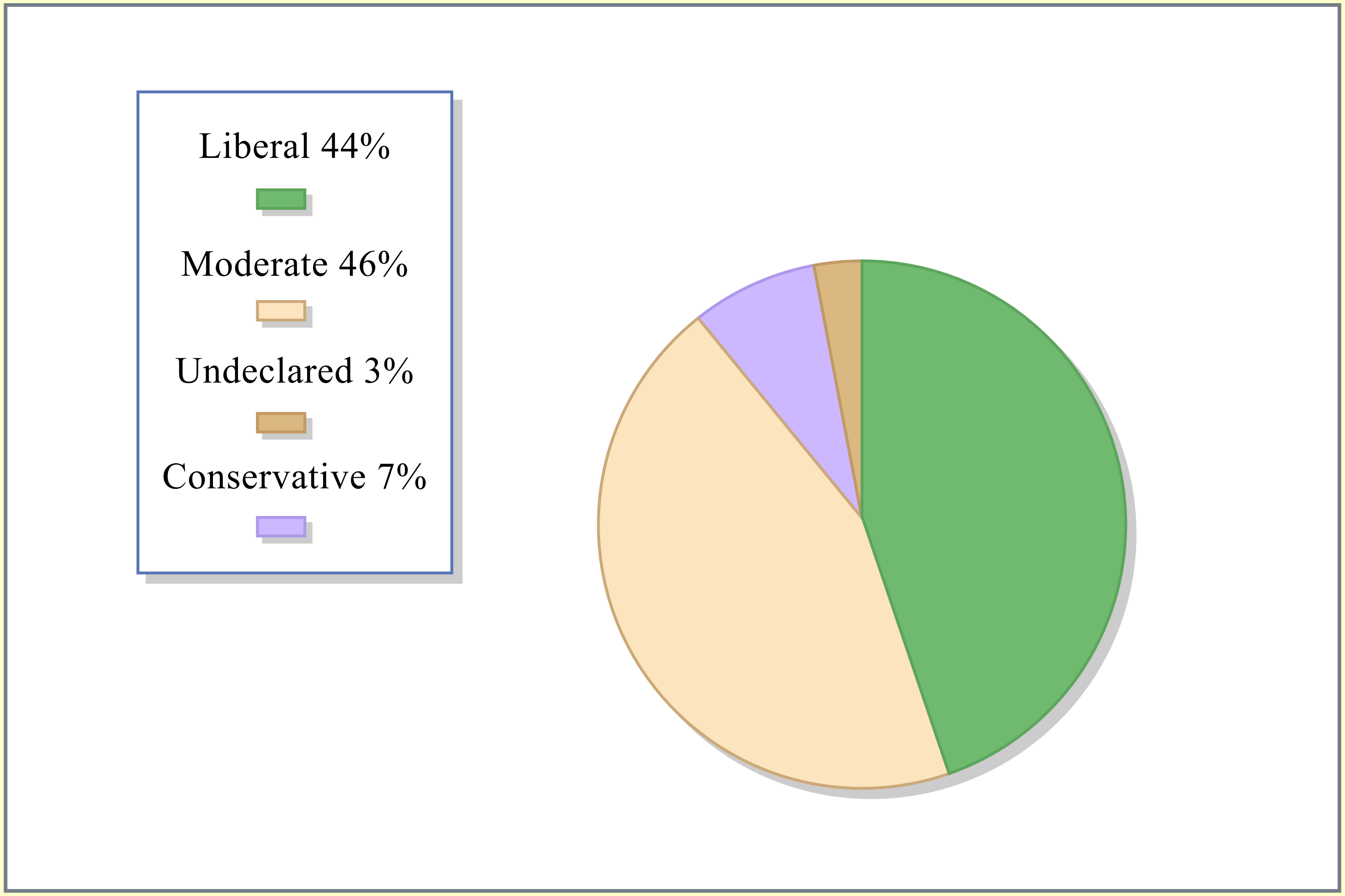


Figure by MIT OCW.

More on Variables

- We often study the effects of one or more explanatory variables on some response variable (or variables).
- We may also study the correlation between two response variables.
  - Is the height of a father correlated with the height of his son?

Cancer Death Rate by Year

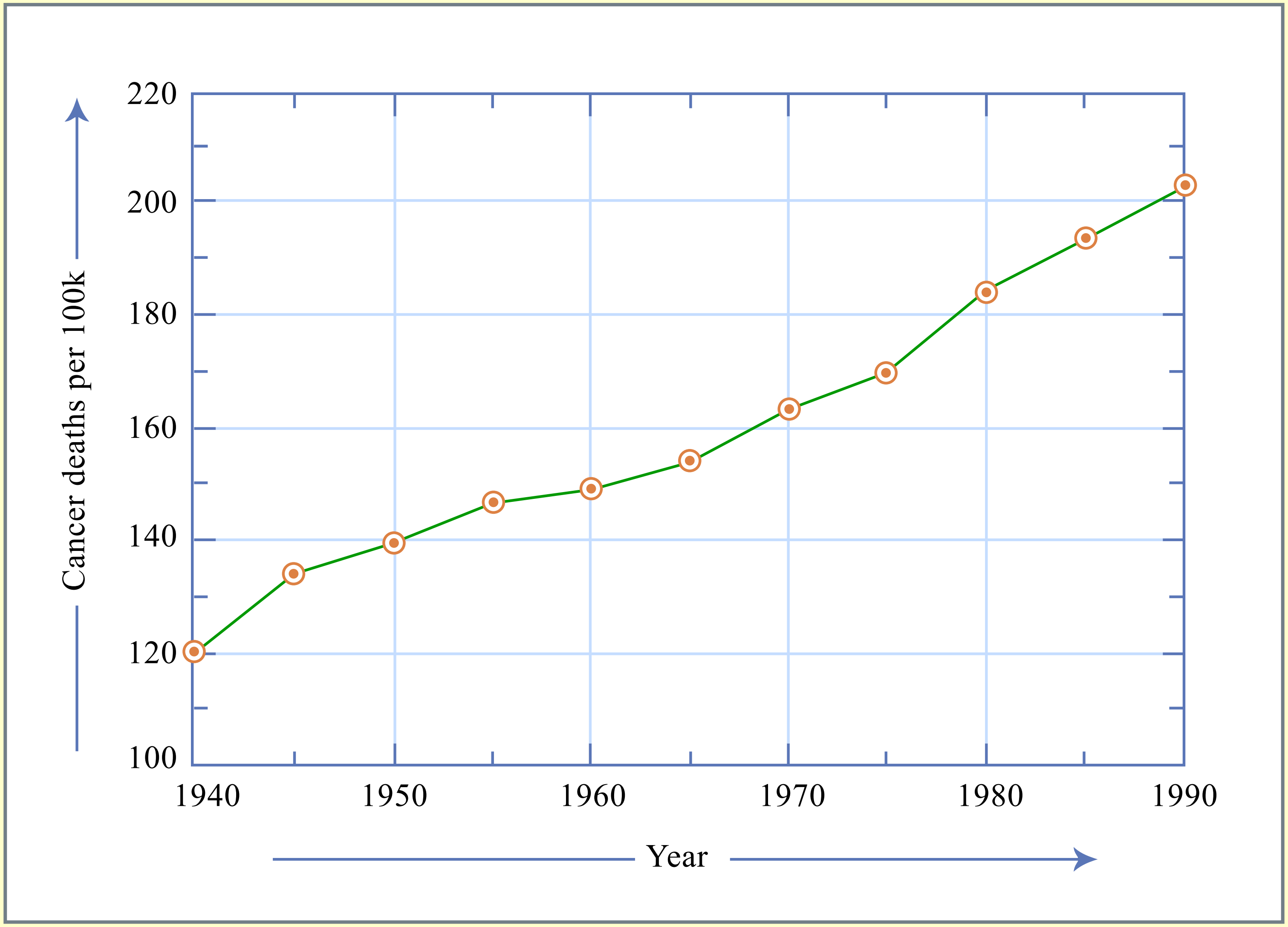


Figure by MIT OCW.

Choosing the Right Graph

- The right form depends on what you're trying to show as well as the type of variable
  - Bar graphs (nominal or ordinal variables) or histograms (interval or ratio variables) to express the sizes of groups
  - Pie charts to express proportions
  - Scatter plots or line graphs to show relationships

Introduction to MATLAB

For MATLAB help

- Type “help” for help on a specific command
  - help randn
- Type “lookfor” if you don’t know the name of the command
  - lookfor histogram
- Ask the TA’s
- Ask the MIT server folks
- Search for a tutorial on the web – they’re all over the place, e.g. <http://www.math.utah.edu/~eyre/computing/matlab-intro/index.html>