- "I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

  Hal Varian , Google Chief Economist, August 2009

- Harvard Business Review (2012)

  Data Scientist: The Sexiest Job of the 21st Century

- New York Times (2009)

  For Today's Graduate, Just One Word: Statistics

- Planning/Design of Study
- Data collection
- Data analysis
- Presentation
- Interpretation

Statistics CAN play a role in most of these steps! (but sometimes is only called upon for the data analysis part)

- Planning/Design of studies
  - Primary Question(s) of Interest:
    - Quantifying information about a single group?
    - Comparing multiple groups?
  - Sample size
    - How many subjects needed total?
    - How many in each of the groups to be compared?
  - Selecting Study Participants
    - Randomly chosen from “master list”?
    - Selected from a pool of interested persons?
    - Take whoever shows up?
  - If group comparison of interest, how to assign to groups?

- Data Collection

- Data Analysis
  - How best to summarize the information coming from the raw data
  - Dealing with variability (both natural and sampling related):
    - Important patterns in data are obscured by variability
    - Distinguish real patterns from random variation
  - Inference: using information from the single study coupled with information about variability to make statement about the larger population/process of interest: What statistical methods are appropriate given the data collected?

- Presentation
  - What summary measures will best convey the “main messages” in the data about the primary (and secondary) research questions of interest
  - How to convey/ rectify uncertainty in estimates based on the data

- Interpretation
  - What do the results mean in terms of practice, the program, the population etc..?

- Throughout all of our endeavors the focus will be on
  - interpreting the results of statistical procedures correctly
  - summarizing the results from published studies in an understandable fashion
  - assessing the strengths and weaknesses of published research results including:
    - study design
    - clarity of the research question(s)
    - appropriateness of the statistical methods
    - clarity of the reported results
    - appropriateness of the overall scientific/substantive conclusions

Learning Objectives

- Upon completing this lecture section you should be able to:
  - Explain the difference between a population and sample (so far as the terms are used in research)
  - Give examples of populations, and of a corresponding sample from a population
  - Explain that characteristics of a randomly selected data sample should imperfectly mimic the characteristics of the population from which the sample was taken
  - Explain how non-random samples may differ systematically from the populations from which they were taken

Populations and Samples

- *Population:* The entire group for which information is wanted
  - For example, all 18-year-old male college students in the United States

- *Sample :* A subset (part) of a larger group (population) from which information is collected to learn about the larger group
  - For example, twenty-five 18-year-old male college students in the United States

Random Sampling

- For studies it is optimal if the sample which provides the data is representative of the population under study
  - Certainly not always possible!
- For this term, we will make this assumption unless otherwise specified
- One way of getting a representative sample: simple random sampling
  - A sampling scheme in which every possible subsample of size *n* from a population is equally likely to be selected

- Generally speaking, with research we want to learn truths in a population, but can only estimate these from an imperfect sample of observations from the population

Example: Sample versus Population, SBP data

- Researchers wanted to learn about the pulmonary health in clinical population of men. There were able to sample 113 men from this population, and measure the systolic blood pressure of each male in the sample.

Example: Sample versus Population, Maternal HIV Transmission Data

- Researchers wanted to characterize the risk of mother to infant HIV transmission (within 18 months of birth). The researchers studied 183 births to HIV+ women and found that 40 of the children tested positive for HIV within 18 months, for a transmission percentage of 22%

Example: Sample versus Population, Pennsylvania Lung Cancer Cases

- Researchers want to study geographic variation in lung cancer cases and potential associated factors (sex distributions, environmental exposures, access to healthcare) using data from a single year for a single US state

Other Types of Non-Random Samples-1

- Other types of sampling may be necessary, but may also result in samples whose elements do not reflect the makeup of the populations of interest (bias)
  - Voters (not registered, but those who will actually vote) in the US Presidential Election
  - Intravenous Drug users in Chennai
  - Patients with a Certain Disease
  - Homeless persons in Baltimore
  - Men who have sex with men (MSM) in Malawi

- Other types of sampling may be necessary, but may also result in samples whose elements do not reflect he makeup of the populations of interest

- What kinds of sampling strategies can be employed that may/may not result in a random sample?

- Voters (not registered, but those who will actually vote) in the US Presidential Election

    - Random digit dialing

- What kinds of other sampling strategies can be employed?
  - Intravenous Drug users in Chennai
  - Homeless persons in Baltimore
  - Men who have sex with men (MSM) in Malawi
    - Convenience Sampling
    - Respondent Driven Sampling

Summary

- Generally speaking, with regards to pubic health and medical research, not all elements of a population can be studied. As such, a sample is taken from the population of interest.
  - Random sampling is the best strategy for getting a sample whose characteristics imperfectly mimic the population
  - However, random sampling is not always feasible: other approaches can be used, and the sampling procedure needs to be considered when applying the results from the sample to the population

Learning Objectives

- At the end of this lecture section you will be able to:
  - Describe the similarities and differences between the randomized cohort, observational cohort, and case-control study designs
  - Explain the major analytical challenge that comes from comparing outcomes across groups where the group membership has not been randomized
  - Start to become aware of some of the major issues to consider when making conclusions based on study results (ie: mapping the statistics to the scientific/clinical/substantive)

Common Study Design Types

- Prospective Cohort Studies
  - Randomized/controlled study design
  - Observational (Cohort) Studies

  Subjects are classified as to their exposure(s) status at study start, and followed over time to see who develops outcome(s)

- Case/Control Studies

  Subjects are chosen based on their outcome status, and the exposure(s) that occurred prior to outcome are assessed

Group Comparisons via Prospective Cohort Studies

- For a randomized study:
  - Get a representative sample from the general population under study (A)
  - Randomly assign sample members to exposure groups

- For an observational study:
  - Get a representative sample from the general population under study (A), and then ascertain group membership
  - Get representative samples from the different populations to be compared

Prospective Cohort Studies: Randomized Trials/Experiments

- Important for accounting for many kinds of biases
- Randomization, done correctly on a large number of subjects nearly ensures that the only systematic difference in the groups being compared is the exposure(s) of interest

Example, Randomized Trial: Salk Polio Vaccine Trial-1

- A very famous randomized trial

200, 745 Vaccinated for Polio

≈ 400,000 School Children Randomized

201,229 Given a Placebo

Example, Randomized Trial: Salk Polio Vaccine Trial-2

- At the end of the follow-up period there were 82 cases in the vaccine group and 162 in the placebo group
- Subsequent analyses report slightly different numbers because some false positives were discovered in each of the two groups

Benefit of Randomization

- Randomization helps protect against self selection biases
  - Examples of such biases
    - Males more likely to volunteer for placebo than females
    - Smokers less likely to be in exposed group
    - Healthier persons sign up for the intervention

- The goal of randomization is to eliminate any systematic differences in characteristics of subjects in each of the exposure groups under study, save for the exposure itself

- Randomization helps protect against self selection biases
  - Examples of such biases
    - Males more likely to volunteer for placebo than females
    - Smokers less likely to be in exposed group
    - Healthier persons sign up for the intervention

- The goal of randomization is to eliminate any systematic differences in characteristics of subjects in each of the exposure groups under study, save for the exposure itself

- Unfortunately (at least for scientific purposes), you cannot always perform randomized trials!!

Smokers

Random Assignment

Non-smokers

- Observational studies are studies in which subjects "self-select" to be in exposure groups: i.e. subjects are not randomized. Sometimes this is the only type of study that can be done

- Outcome/exposure relationships are of interest
  - Sometimes difficult to directly assess because of selection bias issues which may lead to systematic differences between the exposure groups other then the exposure of interest
  - Examples:
    - Smokers more likely to drink alcohol
    - Vegetarians more likely to exercise.

- New York City: relative risk of HIV infection for intravenous drug users (IVDUS) by needle exchange program participation

As per the authors:

" Interpretation :We observed an individual-level protective effect against HIV infection associated with participation in a syringe-exchange programme."

- This was after researchers had accounted for differences (adjusted for) in the program participants and non-participants, including age, gender, race, frequency of injection etc..

Example, Observational Cohort Study: HPV Vaccination and Sexual Activity in Teens

- From the article abstract:

  "**RESULTS:** The cohort included 1398 girls (493 HPV vaccine–exposed;905 HPV vaccine–unexposed)."

  "**CONCLUSIONS:** HPV vaccination in the recommended ages was not Associated with increased sexual activity–related outcome rates."

- The authors made this conclusion after the association was adjusted for other characteristics of the teens including "health care–seeking behavior and demographic characteristics."

Challenges with Regard to Analyzing Observational Studies

- Potential Confounders: confounders are factors that are related to both the outcome and exposure of interest: ignoring these can distort or negate the association of interest
  - Ex: the association between colds and alcohol consumption: those who drink more alcohol may be more likely to smoke cigarettes, and smoking is associated with increase cold risk
- Associations of interest can be adjusted for potential confounders. The nagging question, however, is "what confounders have not been addressed"?

Sometimes, Observational Studies Generate Ideas That Can Be Tested by a Randomized Trial: Beta Carotene and Health

- "**Abstract/Background.** Observational studies suggest that people who consume more fruits and vegetables containing beta carotene have somewhat lower risks of cancer and cardiovascular disease, and earlier basic research suggested plausible mechanisms. Because large randomized trials of long duration were necessary to test this hypothesis directly, we conducted a trial of beta carotene supplementation."

- "**Conclusions.** In this *(randomized)* trial among healthy men, 12 years of supplementation with beta carotene produced neither benefit nor harm in terms of the incidence of malignant neoplasms, cardiovascular disease, or death from all causes.

Case/Control Studies-1

- In the previously discussed prospective cohort-studies (randomized and observational), the subjects had their exposure status assigned to them, or were selected and then the exposure status was classified: the outcome of interest was assessed over time, after the exposure had occurred

- In situations in which researchers wish to study exposures associated with rare outcomes, it is not necessarily feasible to do a prospective cohort study. Such an approach would require a very large number of enrollees in order to see any outcomes in the samples being compared

Example, Case/Control Study: Doll and Hill, Smoking and Lung Cancer-1

- Another landmark public health finding! Subjects were chosen for participation in the study as follows:
  - The method of investigation was as follows: twenty London hospitals were asked to co-operate by notifying all patients admitted to them with carcinoma of the lung, ..” (and several other cancers)

“……for each lung-carcinoma patient visited at a hospital the almoners were instructed to interview a patient of the same sex, within the same five-year age group.”

Example, Case/Control Study: Doll and Hill, Smoking and Lung Cancer-2

- Summary of findings from the article:

  "Consideration has been given to the possibility that the results could have been produced by the selection of an unsuitable group of control patients, by patients with respiratory disease exaggerating their smoking habits, or by bias on the part of the interviewers. Reasons are given for excluding all these possibilities, and it is concluded that *smoking is an important factor in the cause of carcinoma of the lung*."

Doll R and Hill A. Smoking and Carcinoma of the Lung: Preliminary Report, (1950). *British Medical Journal*. pps 739-748.

Challenges with Regard to Analyzing Case/Control Studies

- Potential Confounders: confounders are factors that are related to both the outcome and exposure of interest: ignoring these can distort or negate the association of interest
- Associations of interest can be adjusted for potential confounders. The nagging question, however, is "what confounders have not been addressed"?
- Recall bias
  - Cases and controls may recall exposures differently
  - Exposures assessed after they have occurred, sometimes a very long time after the occurrence: respondent memory can be an issue as well

Learning Objectives

- In this short lecture, a brief summary is given of the types of data that frequently occur in research studies, and will be dealt with analytically in this class (both terms)

- At the end of this lecture section, you should be able to:
  - Distinguish between continuous, binary (and categorical) and time to event data
  - Give examples of each of these aforementioned data types

Continuous Data

- Continuous Data *(incremental measurements)*
  - Blood pressure, mmHg
  - Weight, lbs (kgs, oz etc..)
  - Height, ft (cm, in etc..)
  - Age, years (months)
  - Income level, dollars/year (Euro by year, etc..)

- A defining characteristic of continuous data is that a one unit change in the value means the same thing across the entire range of data values

Binary Data

- Binary (dichotomous) data: takes on only two values, “yes” or “no”

- Binary (dichotomous) data (“Yes/no” data)
  - Polio :Yes/No
  - Remission :Yes/No
  - Sex : Male/Female (or as yes/no, “is subject male?”)
  - Quit Smoking: Yes/No
  - Etc..

Categorical Data

- Categorical data *:* an extension of binary data to include more than 2 possible values

- Nominal categorical data: no inherent order to categories
  - Race/ethnicity
  - Country of birth
  - Religious Affiliation

- Ordinal categorical data: order to categories
  - Income level categorized into four categories, least to greatest
  - Degree of agreement, five categories from strongly disagree to strongly agree

Time-to-Event Data: Two "Formats"

- Count data collected over a fixed period of time
  - Total lung cancer cases occurring in a given year
  - Number of flu diagnoses per week in a given month

- Data that are a hybrid of continuous data and binary data: whether an event occurs and time to the occurrence (or time to last follow-up without occurrence)
  - Time to relapse after remission
  - Time to quitting smoking after treatment

Difference Analysis Tools for Different Data Types-1

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could:
  - Estimate the mean difference in blood pressure change (after-before) between the two treatment groups
  - Estimate a 95% confidence interval and/or use a t-test to test for population level differences in the mean blood pressure change

Difference Analysis Tools for Different Data Types-2

- To compare the proportion of polio cases in the two treatment arms of the Salk Polio vaccine, you could:
  - Estimate the difference in proportions (risk difference) and ratio of proportions (relative risk, risk ratio)
  - Estimate 95% confidence intervals and/or use a chi-square test to test for population level differences in these quantities

Difference Analysis Tools for Different Data Types-1

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could:
  - Estimate the mean difference in blood pressure change (after-before) between the two treatment groups
  - Estimate a 95% confidence interval and/or use a t-test to test for population level differences in the mean blood pressure change

Difference Analysis Tools for Different Data Types-3

- To compare differences in time to contracting HIV between HIV negative IV drug users in a needle exchange program and HIV negative IV drug users not enrolled in a needle exchange program, you could:
  - Estimate an incidence rate ratio for contracting HIV that compares these two groups
  - Construct a Kaplan-Meier curve for each group to provide a graphical description of the time to HIV profile for each group
  - Estimate a 95% confidence interval for the incidence rate ratio and/or use a log-rank to test for a population level

Summary

- The three major types of data we will deal with in the class (and that are of general interest in public health studies) include:
  - Continuous
  - Binary
  - Time-to-event

- There are different approaches for summarizing and analyzing these different data types

Example: Neighborhoods and Health Indicators-1

- First sentence of the article

  "Many observational studies have shown that neighborhood attributes such as poverty and racial segregation are associated with increased risks of obesity and diabetes ,even after adjustment for observed individual and family-related factors.

Example: Neighborhoods and Health Indicators-2

- Additional text includes:

- "It is unclear whether neighborhood environments directly contribute to the development of obesity and diabetes. People living in neighborhoods with high poverty rates differ in many ways from those living in neighborhoods with lower poverty rates, only some of which can be adequately measured in observational studies. These unmeasured individual characteristics may be responsible for variations in health among different neighborhoods."

Example: Neighborhoods and Health Indicators-3

- Moving to Opportunity: randomization period 1994-98
  - Eligible participants: "Families with children (defined as family members younger than 18 years of age) living in Baltimore, Boston, Chicago, Los Angeles, or New York in selected public housing developments in census tracts with poverty rates of 40% or more in 1990 were eligible."
  - Eligible participants were randomized to one of thee groups:
    - Rent vouchers for private market housing, required to be used in a census tract with a low poverty rate (<10% in 1990).
    - Rent vouchers for private market housing, with no restrictions on usage.
    - No assistance.

Example: Neighborhoods and Health Indicators-4

- Outcomes of interest:
  - ”From 2008 through 2010, as part of a long-term follow-up survey, we measured data indicating health outcomes, including height, weight, and level of glycated hemoglobin (HbA1c).”

Example: Neighborhood Disadvantage and CVD-1

- Objective of the research, as per the authors:
  "To examine the impact of neighborhood conditions resulting from racial residential segregation on cardiovascular disease (CVD) risk in a socioeconomically diverse African American sample."

Example: Neighborhood Disadvantage and CVD-2

- Study Sample and Exposure
  - "The study included 4096 African American women (n = 2652) and men(n = 1444) aged 21 to 93 years from the Jackson Heart Study (Jackson, Mississippi; 2000–2011). We assessed neighborhood disadvantage with a composite measure of 8 indicators from the 2000 US Census. We assessed neighborhood-level social conditions, including social cohesion, violence, and disorder, with self-reported, validated scales"
  - "The analytic sample included all participants with geocoded information who resided in the Jackson metropolitan area and were free of CVD at baseline (n = 4698)."

Example: Neighborhood Disadvantage and CVD-3

- Primary Outcome
  - Incidence of CVD during the study follow-up period
  - "We followed participants with geocoded information free of CVD at baseline (n = 4968) from the time of their baseline examination in 2000 to 2004 to the date of their first CVD event, death, and loss to follow-up, or otherwise through December31, 2011.

1. Which is the following would be classified as a continuous data measure?

- ○ Gender Identity (cis male, cis female, non-binary, trans, other)
- ○ Disease status: disease or no disease
- ◉ Expenditures (in US dollars) incurred during last inpatient hospital visit
- ○ Degree of agreement on a 5 point scale (strongly disagree, disagree, neutral, agree, strongly agree)

2. A randomized controlled clinical trial is an example of a(n):

- ○ cross-sectional study
- ○ observational cohort study
- ◉ prospective cohort study
- ○ case-control study

3. A study is designed to assess the relationship between a rare disease ("disease A") and an exposure of interest ("exposure A"). One-hundred subjects with "disease A" are recruited to participate in the study, as are 200 subjects who do not have "disease A". This is an example of what kind of study design?

- (●) Case-control study
- ( ) Randomized prospective cohort study
- ( ) Cross-sectional study
- ( ) Observational prospective cohort study

4. What is the primary advantage of a randomized controlled trial (RCT) as compared to other study designs?

- ( ) RCTs tend to be less expensive than other types of studies.
- ( ) Researchers who do RCTs are better scientists than those who perform other types of studies.
- ( ) The results from a RCT are likely to be more interesting scientifically as compared to results from the other types of studies.
- (●) The likelihood that the outcome/exposure relationship is confounded by other factors is minimized with a RCT design.

5. Which of the following is a characteristic of a random (representative) sample taken from a larger population?

- ○ These samples tend to be smaller than non-random samples.
- ○ The characteristics of the sample will be exactly the same as the characteristics of the population from which the sample is taken.
- ◉ The characteristics of the sample should be similar to the characteristics of the population from which the sample is taken.
- ○ Random samples are less expensive to obtain than non-random samples.

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Public Health Statistics: Continuous Data Measures

John McGready, PhD
Johns Hopkins University

Learning Objectives

- Upon completion of this lecture, you will be able to:
  - Compute a sample mean and standard deviation
  - Interpret the estimated mean, standard deviation, median, and various percentiles computed for a sample of continuous data measures

Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median (50th percentile)
- Measure of data variability
  - Standard deviation
- Other measures of location
  - Percentiles

Sample Mean: The Average or Arithmetic Mean

- Add up data, then divide by sample size (n)
- The sample size n is the number of observations (pieces of data)
- Example: mean, small systolic blood pressure (SBP) dataset

Example: Mean, Small SBP Dataset—2

- Five systolic blood pressures (mmHg), n=5: 120 mmHg, 80 mmHg, 90 mmHg, 110 mmHg, 95 mmHg
- Can be represented with math type notation: x_1= 120, x_2=80,...x_5=95
- The sample mean:

$$\bar{x} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99\ mmHg$$

The Sample Mean, Generally Speaking—1

- Generic formula representation

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \text{where} \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots x_n$$

- In the formula, to find the mean, we use the "summation sign," $\sum$
  - This is just mathematical shorthand for "add up all of the observations"

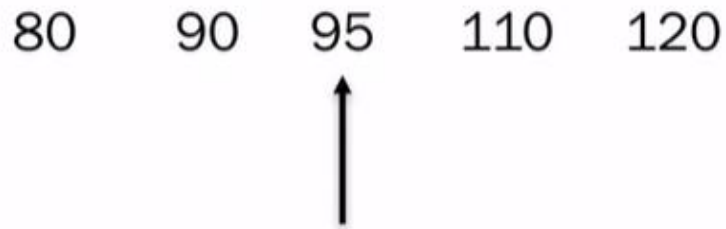
- Also called *sample average* or *arithmetic mean*

- Why is it called the *sample* mean?
  - To distinguish it from population mean (an unknown, unknowable value of interest μ, that can be estimated by $\bar{x}$)

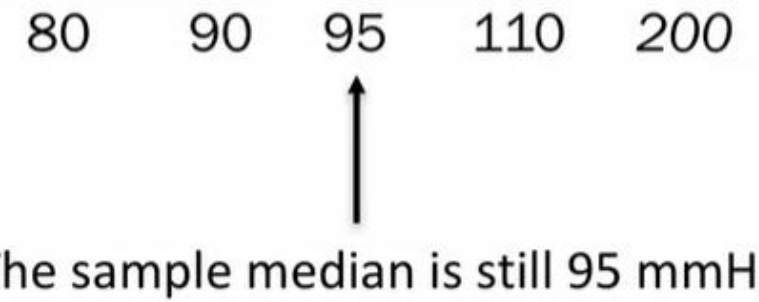
- Sensitive to extreme values (in smaller samples)
  - A change in the value of one data point could make a substantial change in the values of a sample mean

- The median is the middle value in an ordered set of continuous data measures (the median is also called the 50^{th} percentile)

- The median value of the five SBP measurements is 95 mmHg:



- The sample median is not sensitive to the influence of extreme sample values (unlike the sample mean)
  - For example, in the sample of five SBP measurements, if the value 120 was changed to 200, the sample median would remain the same, but the *sample mean would increase from 99 mmHg to 115 mmHg*



Example: The Median, Small SBP Data Set—3

- If the sample size is an even number, then the median is the average of two middle values
- Suppose we add a sixth SBP value to the original SBP values. The (now) six values, in ascending order:

80 90 95 110 120 125

The sample median of these six values is $\frac{95+110}{2} = 102.5$ mmHg

Describing Variability: The Sample Variance

- Sample variance (s^2)
- Sample standard deviation (s or SD)
- The sample variance is the average of the square of the deviations about the sample mean

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

Describing Variability: Sample Standard Deviation (s)

- The sample standard deviation is the square root of the sample variance, s^2

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

Example: Standard Deviation, Small SBP Data Set—1

- Systolic blood pressures (mmHg), n=5: 120 mmHg, 80 mmHg, 90 mmHg, 110 mmHg, 95 mmHg. The mean, $\bar{x}$, is 99 mmHg.
- The sample variance computation, numerator:

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 =$$

$$\sum_{i=1}^{5}(x_i - 99)^2 = 21^2 + (-19)^2 + (-9)^2 + 11^2 + (-4)^2 = 1{,}020 \text{ mmHg}^2$$

The Sample Standard Deviation (s), Generally Speaking—1

- The more variability there is in a sample of data, the larger the value of s
- s measures the variability (spread) of the individual sample values around the sample mean
- s can equal 0 only if there is no variability (if all n sample observations have the same value)
- The units of s are the same as the units of the data measurements in the sample (for example, mmHg)
- Often abbreviated *SD* or *sd*
- s^2 is the best estimate from the sample of the population variance σ^2; *s is the best estimate of the population standard deviation* σ

Example: Standard Deviation, Larger SBP Data Set—1

- SBP measurements taken for a sample of 113 men (n=113)
- The first 50 sample values are shown here:

142	116	137	126	124
123	116	127	115	129
107	103	130	133	116
129	117	131	107	138
114	113	142	120	147
105	122	111	111	129
132	89	134	121	120
128	120	119	112	139
121	124	132	140	120
116	152	123	131	141

Percentiles in Samples With All Unique Values

- Other values that can help us quantify the distribution of continuous data values include the sample percentiles (as estimates of the underlying population percentiles)
- In general, if all sample values are unique, the p^{th} sample percentile is that value in a sample of data such that p percent of the sample values are less than or equal to this value, and $(100-p)$ percent are greater than this value (example: the median is the 50th percentile)
- Percentiles can be computed by hand but are generally done via computer

Example: Percentiles, Larger SBP Data Set—1

- Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population (based on results from a computer)
- The *10th percentile* for these 113 blood pressure measurements is 107 mmHg, meaning that approximately 10% of the men in the sample have SBP ≤ 107 mmHg, and (100−10) = 90% of the men have SBP > 107 mmHg
- The *75th percentile* for these 113 blood pressure measurements is 132 mmHg, meaning that approximately 75% of the men in the sample have SBP ≤ 132 mmHg, and (100−75) = 25% of the men have SBP > 132 mmHg

Summary

- Summary measures that can be computed on a sample of continuous data include the mean, standard deviation, median (50^{th} percentile), and other percentiles

- These sample-based estimates are the best estimates of unknown, underlying population quantities. For example:
  - $\bar{x}$ is the best estimate of the population mean (μ)
  - s is the best estimate of the population standard deviation (σ)

- (Soon) we will discuss how to address the uncertainty in the estimates of certain sample quantities (ex: mean)

Pictures of Data: Continuous Variables

- Histograms and boxplots
  - Means, standard deviations, and percentile values do not tell the whole story of data distributions
  - Differences in shape of the distribution
  - Histograms are a way of displaying the distribution of a set of data by charting the number (or percentage) of observations whose values fall within pre-defined numerical ranges
  - Boxplots are graphics that display key characteristics of a dataset: these are especially nice tools for comparing data from multiple samples visually

Example: Histogram, 113 Systolic Blood Pressures

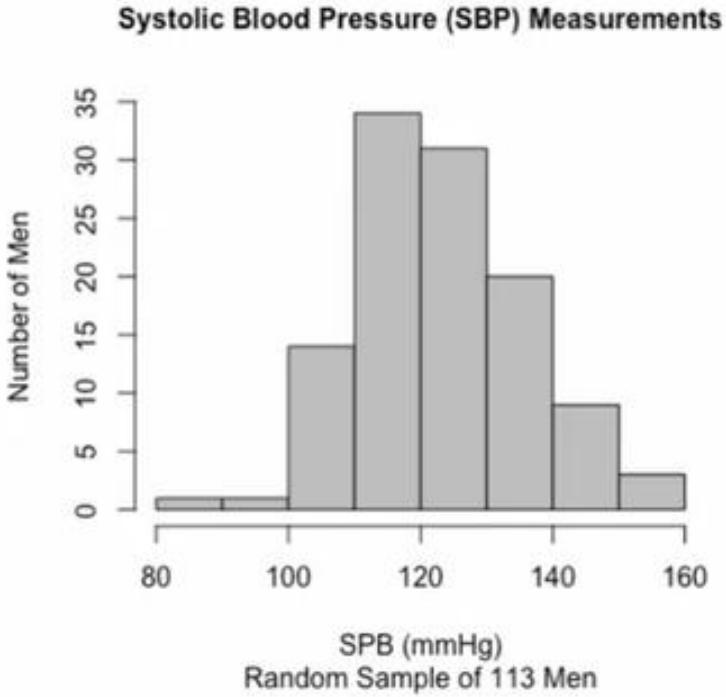
- Data on systolic blood pressures (SBP) from a random clinical sample of 113 men

- A histogram can be created by:
  - Breaking the data (blood pressure) range into bins of equal width
  - Counting the number of the 113 observations whose blood pressure values fall within each bin
  - Plotting the number (or relative frequency) of observations that fall within each bin as a bar graph

Example: Histogram, 113 Systolic Blood Pressures: Number of Observations

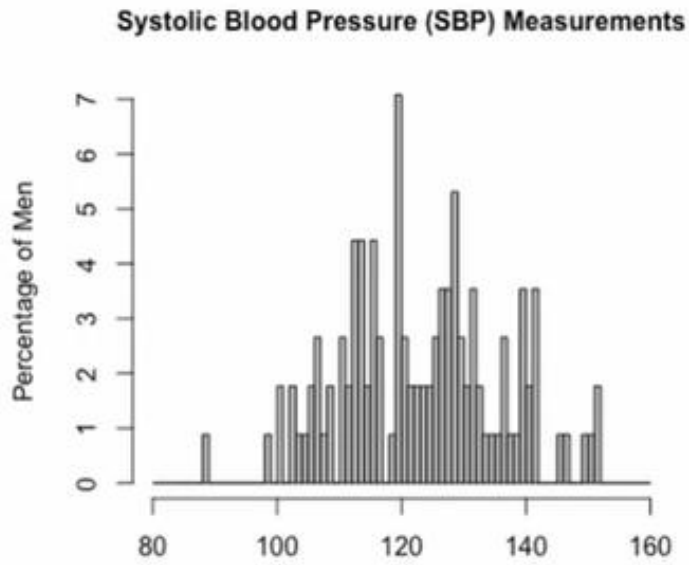
- A basic histogram of these 113 measurements
  - Number of observations on the vertical axis

$$\bar{x} = 123.6 \text{ mmHg}$$
$$s = 12.9 \text{ mmHg}$$
$$\hat{m} = 123.0 \text{ mmHg}$$



Example: Histogram, 113 Systolic Blood Pressures: Narrower Bin Width

- A basic histogram of these 113 measurements
  - Percentage of observations on the vertical axis, narrower bin width



Summary

- Histograms and boxplots are useful visuals tools for characterizing the shape of a data distribution above and beyond the information given by summary statistics
- Relatively common shapes for samples of continuous data measures include symmetric and “bell” shaped, right skewed, left skewed, and uniform

Learning Objectives

- Upon completion of this lecture section, you will be able to:
  - Understand that a random sample taken from a larger population will (imperfectly) mimic the characteristics of the larger population
  - Understand that the distribution of values in a random sample should reflect the distribution of the values in the population from which the sample was taken
  - Understand and explain that increasing sample size does not systematically decrease the value of sample summary statistic estimates
  - Begin to understand that while increasing sample size does not decrease sample summary statistic estimates, the estimates become less variable with larger samples

Summary

- The distribution of sample values of continuous data should (imperfectly) mimic the distribution of the values in the population from which the sample was taken

- With regard to the distribution of sample values and increasing sample size:
  - Will not systematically alter the shape of the sample distribution
  - Will result in a more "filled out" distribution
  - Will not systematically alter the values of the sample statistic
    - The sample statistic estimates will vary from random sample to random sample but will not systematically get larger (or smaller) with increasing sample size
  - Will increase the precision of the summary statistics as estimates of the unknown (population level) true values (more to come shortly)

Learning Objectives

- Upon completion of this lecture section, you will be able to:
  - Suggest graphical approaches to comparing distributions of continuous data between two or more samples
  - Explain why a difference in sample means can be used to quantify, in a single number summary, differences in distributions of continuous data

Motivation: Comparisons

- Frequently, in public health/medicine/science, etc., researchers/practitioners are interested in comparing two (or more) populations via data collected on samples from these populations

- Such comparisons can be used to investigate questions, such as:
  - How does weight change differ between those who are on a low-fat diet compared to those on a low-carbohydrate diet?
  - How do salaries differ between males and females?
  - How do cholesterol levels differ across weight groups?

- While these comparisons can be done visually, it is also useful to have a numerical summary

Motivation: Comparisons

- Frequently, in public health/medicine/science, etc., researchers/practitioners are interested in comparing two (or more) populations via data collected on samples from these populations

- Such comparisons can be used to investigate questions, such as:
  - How does weight change differ between those who are on a low-fat diet compared to those on a low-carbohydrate diet?
  - How do salaries differ between males and females?
  - How do cholesterol levels differ across weight groups?

- While these comparisons can be done visually, it is also useful to have a numerical summary

Motivation: Numerical Summary

- Theoretically, this numerical summary could be many things:
  - Difference in medians
  - Ratio of means
  - Difference in 95^{th} percentiles
  - Ratio of standard deviations
  - Etc.

- However, what is commonly used (for reasons that we will elaborate on shortly) is a difference in sample means
  - When comparing sample distributions, this can be a reasonable measure of the overall differences in these distributions (as an estimate of the underlying difference in the population distributions)

Example: Academic Physician Salaries: Comparing Means between More than Two Groups

- General practice for numerically comparing means between more than two groups
  - Designate one group as the "reference"
  - Report the differences for the other groups compared to this same reference

- For example, if we make "West" the reference region, then the mean differences can be reported as:

$$\bar{x}_{MW} - \bar{x}_W = 198{,}890 - 194{,}474 = \$4{,}416$$
$$\bar{x}_S - \bar{x}_W = 194{,}439 - 194{,}474 = -\$35$$
$$\bar{x}_{NE} - \bar{x}_W = 192{,}152 - 194{,}474 = -\$2{,}322$$