against all odds

inside statistics

- **statistics** helps with data organization to identify if patterns are truly patterns
- **statistics** is used in all aspects of life
  - **descriptive statistics** demonstrates stats in an informative way
  - **inferential stats** comes to conclusions based on a sample
- **probability**: mathematical way to assess chance of events

- **lateral reading**: stat checking on other websites to ensure accuracy online
  - look at authority and perspective

authority ↓

knowledge /expertise
leaving sites to fact check; not vertical

↓

look at process of gathering info, systems to catch mistakes, professional background

↓

when mistakes are made, reliable sources issue a statement of correction

perspective →

bias / point of view, opinion piece vs. article

historical notes

1. in the beginning, stats involved charts and tables

2. the chinese used stats for keeping state records and warrior availability ; 2000 BC

3. John Graunt began the study of statistics in 1662

4. statistic theory wasn't commonly used prior to the 1930's because the accumulation and analysis of statistical data involved time-consuming and complicated calculations. changed with the invention of computers

5. inference: making generalizations on the basis of samples

6. the origins on the study of probability are found in correspondences between Blaise Pascal and Pierre Fermat in the 1600s in France

7. Girolomo Cardan; wrote the Book on Games of Chance, which was a book on the theory of randomness

Historical notes

Although statistics is one of the oldest branches of mathematics, it was not until the twentieth century that its use became widespread. Originally, it involved summarizing data by means of charts and tables. Historically, the use of statistics can be traced back to the ancient Egyptians and Chinese who used statistics for keeping state records. The Chinese under the Chou Dynasty, 2000 B.C., maintained extensive lists of revenue collection and government expenditures. They also maintained records on the availability of warriors.

The study of statistics was really begun by an Englishman John Graunt (1620–1674). In 1662 he published his book, *Natural and Political Observations Upon the Bills of Mortality*. Graunt studied the causes of death in different cities and noticed that the percentage of deaths from different causes was about the same and did not change considerably from year to year. For example, deaths from suicide, accidents, and certain diseases not only occurred with surprising regularity but with approximately the same percentage from year to year. Furthermore, Graunt's statistical analysis led him to discover that there were more male than female births. But, since men were more subject to death from occupational hazards, diseases, and war, it turned out that at marriageable age the number of men and women was about equal. Graunt believed that this was nature's way of assuring monogamy.

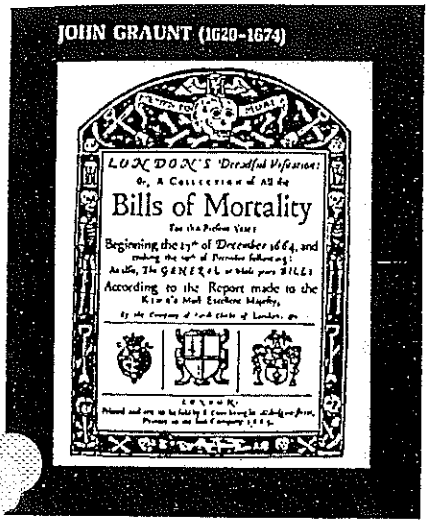


FIGURE 1.4 Illustration "Bills of Mortality" redrawn from *Devils, Drugs, and Doctors* by Howard W. Haggard, M.D. Copyright 1929 by Harper and Row, Publishers, Inc.; renewed 1957 by Howard W. Haggard. Reprinted by permission of the publisher.

After Graunt published his *Bills of Mortality*, many other mathematicians became interested in statistics and made important contributions. Pierre-Simon Laplace (1749–1827), Abraham De Moivre (1667–1754), and Carl Friedrich Gauss (1777–1855) studied and applied the **normal** distribution (see page 270). Karl Pearson (1857–1936) and Sir Francis Galton (1822–1911) studied the correlation coefficient (see page 423). These are but a few of the many mathematicians who made valuable contributions to statistical theory. In later chapters we will further discuss their works.

Although a great deal of modern statistical theory was known before 1930, it was not commonly used, simply because the accumulation and analysis of statistical data involved time consuming, complicated computations. However, things changed with the invention of the computer and its ability to perform long and difficult calculations in a relatively short period of time. Statistics soon began to be used for **inference**, that is, in making generalizations on the basis of samples. Also, probability theory was soon applied to the statistical analysis of data. The use of statistics for inference resulted in the discovery of new techniques for treating data.

Interestingly enough, the principles of the theory of probability were developed in a series of correspondences between Blaise Pascal (1623–1662) and Pierre Fermat (1602–1665). Pascal was asked by the Chevalier de Méré, a French mathematician and professional gambler, to solve the following problem: In what proportion should two players of equal skill divide the stakes remaining on the gambling table if they are forced to stop before finishing the game? Although Pascal and Fermat agreed on the answer, they both gave different proofs. It is in these correspondences during the year 1654 that they established the modern theory of probability.

A century earlier the Italian mathematician and gambler Girolomo Cardan (1501–1576) wrote *The Book On Games Of Chance*. This is really a complete textbook for gamblers since it contains many tips on how to cheat successfully. The origins of the study of probability are to be found in this book. Cardan was also an astrologer. According to legend, he predicted his own death astrologically and to guarantee its accuracy he committed suicide on that day. (Of course, that is the most convincing way to be right!) He also had a temper and is said to have cut off his son's ears in a fit of rage.

Finding Your Way through a Space of Possibilities

In the years leading up to 1576, an oddly attired old man could be found roving with a strange, irregular gait up and down the streets of Rome, shouting occasionally to no one in particular and being listened to by no one at all. He had once been celebrated throughout Europe, a famous astrologer, physician to nobles of the court, chair of medicine at the University of Pavia. He had created enduring inventions, including a forerunner of the combination lock and the universal joint, which is used in automobiles today. He had published 131 books on a wide range of topics in philosophy, medicine, mathematics, and science. In 1576, however, he was a man with a past but no future, living in obscurity and abject poverty. In the late summer of that year he sat at his desk and wrote his final words, an ode to his favorite son, his oldest, who had been executed sixteen years earlier, at age twenty-six. The old man died on September 20, a few days shy of his seventy-fifth birthday. He had outlived two of his three children; at his death his surviving son was employed by the Inquisition as a professional torturer. That plum job was a reward for having given evidence against his father.

Before his death, Gerolamo Cardano burned 170 unpublished manuscripts.[1] Those sifting through his possessions found 111 that survived. One, written decades earlier and, from the looks of it, often

revised, was a treatise of thirty-two short chapters. Titled *The Book on Games of Chance*, it was the first book ever written on the theory of randomness. People had been gambling and coping with other uncertainties for thousands of years. Can I make it across the desert before I die of thirst? Is it dangerous to remain under the cliff while the earth is shaking like this? Does that grin from the cave girl who likes to paint buffaloes on the sides of rocks mean she likes me? Yet until Cardano came along, no one had accomplished a reasoned analysis of the course that games or other uncertain processes take. Cardano's insight into how chance works came embodied in a principle we shall call the law of the sample space. The law of the sample space represented a new idea and a new methodology and has formed the basis of the mathematical description of uncertainty in all the centuries that followed. It is a simple methodology, a laws-of-chance analog of the idea of balancing a checkbook. Yet with this simple method we gain the ability to approach many problems systematically that would otherwise prove almost hopelessly confusing. To illustrate both the use and the power of the law, we shall consider a problem that although easily stated and requiring no advanced mathematics to solve, has probably stumped more people than any other in the history of randomness.

levels of measurement

- nominal data
  - names, labels, listings
  - order not important, no rankings
  - ex. colors, list of names, students
- ordinal data
  - can be arranged in order, no difference between values
  - ex. rankings (low, med, high) ← numerical difference
- interval data
  - can be arranged in order and differences between values are meaningful
  - no natural zero or starting point
  - olympic years, temp., etc
- ratio data
  - highest level of quantitative data
  - has order, meaningful differences, can be divided, has a starting 0
  - ex. distance from somewhere, time it takes to do something

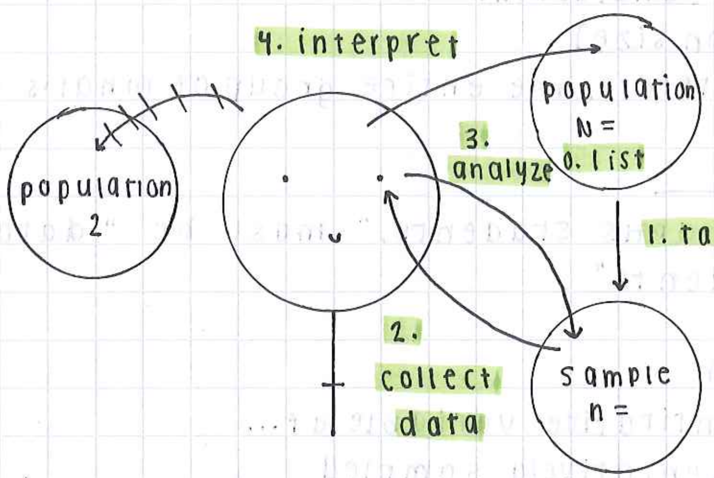
categorical

quantitative

intro to statistics

- statistical data helps make informed decisions when faced with uncertainty without statistical bias
  - collection, organization, analysis, and interpretation of numerical information

↳



↳

collect sample experiment census simulation	organize
analyze	interpret

experiments are the ONLY ones that show causation

census is the least efficient

stats book pg 24/25 for examples:

- anecdotes: investigator recounts instances only known to him/her
- surveys: ask questions
- observational studies: investigator passively observes and records information
  - subject chooses what to do
- experiments: deliberately imposes conditions on experimental units (subjects), observing and recording

- organizing data: frequency tables, charts, graphs
- analyzing: center, shape, spread, outliers, trends (as x increases, what does y do?)
  - central tendency, dispersion, confidence intervals

- interpretation: assumptions based on sample
  - p-value and alpha level
  - confidence levels
  - z-score, T-score, chi-squared scores
  - written summaries and conclusions
- population (N = population size)
  - all measurements of interest, the entire group of what's being measured
  - data from all ___________
    - ↳ cannot just be "all OPHS students," must be "data from all OPHS students"
- sample (n = sample size)
  - part of the population
  - data from... the quantitative variable of...
  - randomly and representatively sampled

- random samples out of complete lists of everyone = fair and equal chance of being selected
  - cannot be chosen out of specific groups where everyone is too similar
  - start with a complete LIST
- inferential statistics (diagram on front)
  - you can only make inferences about the population you sampled from
  - increase confidence by increasing sample size
    - ↳ do it too much and it becomes a census
- limits of inferential stats in article
  - location
  - ethnicity
  - gender
  - age (too young bc parkinsons hits later)
  - observational study
    - ↳ not causation bc not experiment
    - ↳ lifestyle of population sample are different
- obtuse ollie: avg # of women in CA who are university grads who get married for the first time
  - LIST from the marriage license bureau
  - $n = 100$; $\bar{x} = 27.8$ → collect and analyze
  - $\mu \approx 28 \pm 1.5$

- bias: nonresponse, undercoverage, response, convenience sample / selection bias, voluntary response bias, confirmation and availability bias, etc
- conv. sampling disproportionately represents population
  - bias and inaccurate conclusions
- voluntary response bias
  - usually passionately negative or positively responders
- fight the bias!
  - start with a LIST
  - sample
    - ↳ simple random, stratified, cluster, systematic

Sample = caffeine and parkinson's of 8004 Japanese American men in Hawaii

Caffeine May Prevent Parkinson's

By LINDSEY TANNER, AP Medical ...er

CHICAGO--A new study published today suggests that coffee may prevent Parkinson's disease, the degenerative brain disorder that affects more than 1 million Americans.

How a product that makes people jittery could keep them from getting a disease that gives them tremors is not examined in the study of 8,004 Japanese-American men in Hawaii.

But the researchers said the benefits are probably due to caffeine -apparently the more, the better -and they suggest some theories about how it might work.

Outside experts said that if the findings hold up, they could lead to ways to treat Parkinson's more effectively or even prevent the disease.

The study found that men who didn't drink coffee were five times more likely to develop Parkinson's than those who drank the most -4 1/2 to 5 1/2 6 -ounce cups a day. Non-coffee drinkers were two to three times more likely to get the ...se than men who drank 4 ounces to ... cups a day.

The researchers said it is uncertain whether their results would hold true in women and other ethnic groups.

The study was published in today's Journal of the American Medical Association. It was led by Dr. G. Webster Ross, a neurologist at the Veterans Administration Medical Center in Honolulu.

Ross said it is possible that heavy coffee drinkers have a brain composition that may make them resistant to Parkinson's. Previous studies have found low rates of Parkinson's in "thrill-seeking" people who tend to engage in high-risk behavior like engage in high-risk behavior like smoking and heavy drinking, and heavy coffee drinking also fits that personality profile, he said.

But he also suggested that caffeine may somehow protect against the nerve-cell destruction that causes Parkinson's.

Still, Ross said it is too early to recommend coffee as a treatment.

"Hopefully, this will lead to more basic research on caffeine and its effect on areas of the brain affected by Parkinson's disease," Ross said.

Ross said his study was larger than similar previous research and took into account other factors that could explain the findings, such as cigarette smoking, which has also been linked to a decreased Parkinson's risk.

Paul Carvey, director of the neuropharmacology research laboratories at Rush-Presbyterian-St. Luke's Medical Center in Chicago, said the study is important because it traced the benefits to caffeine, showing similar results with caffeine-laden foods other than coffee.

Dr. Abraham Lieberman, medical director for the National Parkinson Foundation, called the results "very interesting and very provocative." He said that if caffeine does have benefits, it is unclear whether it can actually prevent Parkinson's or slow its progression.

Parkinson's is usually associated with aging, though it has made headlines recently with actor Michael J. Fox's disclosure that he was diagnosed seven years ago at age 30. Attorney General Janet Reno and Muhammad Ali are among others with Parkinson's.

The disease involves gradual deterioration of nerve cell clusters that make the chemical dopamine, which helps control muscle movements. Ross and colleagues speculated that caffeine might increase dopamine levels.

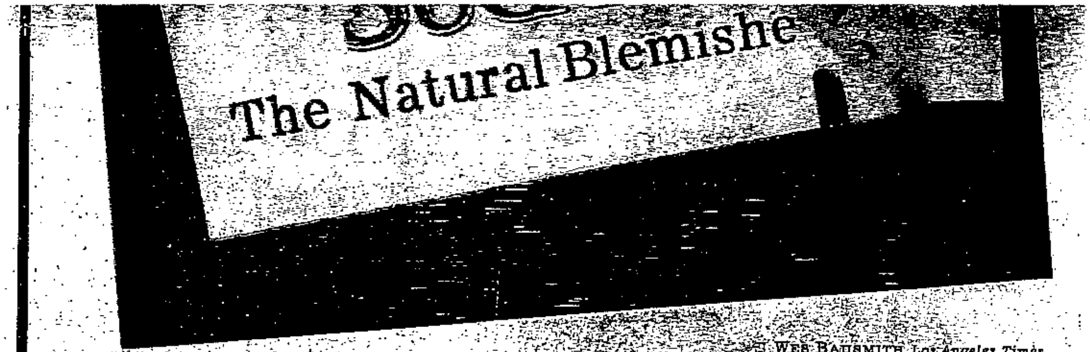
Symptoms of Parkinson's include hand and head tremors, loss of balance, and stiffness. Dementia and depression also can result.

Medication helps victims function, but over time the disease usually renders patients unable to care for themselves. Its cause is unknown.

The researchers examined data from the ongoing Honolulu Heart Program. Participants -age 53 on average when the study began -were asked about coffee consumption at the outset in 1965 and again in 1971. The researchers then measured Parkinson's disease rates from 1991 to 1996. The disease developed in 102 men.

On the Net: National Parkinson's Foundation: http://www.parkinson.org

Parkinson's Disease Foundation: http://www.pdf.org



WES BAUSMITH *Los Angeles Times*

Diet may trigger acne after all

By DIANNE PARTIE LANGE
Special to The Times

FOR years it was widely believed that certain foods, such as chocolate and French fries, made acne worse. Then dermatologists said food didn't cause pimples. Now, get ready for another about-face.

According to a study in the December issue of the Archives of Dermatology, our Western diet may be a reason 79% to 95% of American teenagers have acne. Researchers spent seven weeks examining the skin and lifestyle of village people on Kitava Island, Papua New Guinea. No acne was found in the more than 1,200 people studied, including 300 15- to 25-year olds.

P
D, L, A of
all village ppl
of PNG

S: diet, lifestyle
acne of 1200
village people
in Kitava,
Papua New
Guinea

Unlike typical American teenagers, Kitava islanders are physically active and eat a low-fat (20%), high-carbohydrate (70%) diet of mostly roots, fruits and vegetables, which keeps their insulin levels low. The researchers found no acne in a group of 115 Ache people in Paraguay either. They also eat a low-glycemic diet, but they do eat animal protein.

Studies have shown that when insulin levels in the blood peak, a series of hormonal events increases production of testosterone and several potent growth factors. Testosterone stimulates sebum, or oil, production in the pores. The growth factors cause an overgrowth of cells lining the pores, which creates a plug, keeping the oil in.

"It's like a balloon with no outlet that then becomes infected and causes acne," says Loren Cordain, a coauthor of the study and a specialist in evolutionary medicine at Colorado State University.

High-glycemic foods that increase insulin and are implicated in acne include white flour, sugar and potatoes — ubiquitous in the West, says Cordain. And about that chocolate and French fries? He says it's not the fat that's the problem, it's the sugar.

chapter one

- statistics is the study of how to collect, organize, analyze, and interpret numerical information from data
  - both the science of uncertainty and the technology of extracting information from data
- first you must gather data; identify the individuals or objects to be included in the study and the characteristics or features of the individuals of interest
  - individuals: people or objects involved in the
  - variable: characteristic of the individual to be measured or observed (ex. height, age, weight, gender, etc)
    - ↳ quantitative variable has a value or numerical measurement for which operations such as addition or averaging make sense
    - ↳ qualitative variable describes an individual by placing the individual into a category or group (non-numerical)
- identify the data source
  - population data: every individual of interest
  - sample data: only some individuals of interest
- a population parameter is a numerical measure that describes an aspect of a population
  - ex. data from all individuals who climbed Mt. Everest is population data
    - ↳ population parameter would be proportion of males in the population of all climbers who climbed Mt. everest
- sample statistic is a numerical measure that describes an aspect of a sample
  - proportion of male climbers in the sample is an example of a statistic
    - ↳ different samples may have different values
  - sample stats can vary, population parameters are fixed
- nominal data: names, labels, or categories
  - cannot be organized from smallest to largest
- ordinal data: can be arranged in order but differences between values cannot be determined or are meaningless
- interval data: can be arranged in order, differences between data values are meaningful
- ratio data: can be arranged in order and differences and ratios of data are meaningful
  - set zero exists

- <u>simple random sample</u> of n measurements from a population is a subset of the population selected in such a manner that every sample of size n from the population has an equal and fair chance of being selected
  - every <u>sample has an equal chance</u> and every <u>individual has an equal chance</u>
    - ↳ for a simple random sample, every sample of the given size must also have an equal chance of being selected
  - can be made by numbering individuals and randomly picking a certain amount
    - ↳ also use a random number table
- <u>simulation</u>: a number facsimile or representation of real life
  - process of providing numerical imitations of "real" phenomena
  - productive in studying almost all aspects of modern life
- <u>stratified sampling</u> uses groups or classes inside a population that share a common characteristic
  - <u>strata</u> (groups) are often sampled <u>in proportion</u> to their actual percentages of occurrence in the overall population
- <u>systematic sampling</u>
  - assume the elements of the population are arranged in some natural sequential order
  - select a random starting point and select <u>every x th element</u> for the sample
    - ↳ if population is repetitive or cyclic in nature, don't use
- <u>cluster sampling</u>
  - divide the demographic area into sections and then randomly select sections or <u>clusters</u>
    - ↳ every member of the cluster is included in the sample
- <u>multistage sample design</u> for large or geographically spread out populations
  - select samples of large geographic areas
  - break them down and stratify according to various factors
  - break down even further until clusters are made
    - ↳ <u>final cluster is interviewed</u>
- <u>convenience samples</u> use results or data that are conveniently
  - very risky for severe <u>bias</u>

- a sampling frame is a list of individuals from which a sample is actually selected
  - not all members may be accessible
- undercoverage results from omitting population members from the sample frame
  - when the sample frame doesn't match the population
  - homeless, fugitive, etc might not be included in demographic studies
- sampling error: difference between measurements from a sample and corresponding measurements from the respective population
  - caused by the fact that the sample doesn't perfectly represent the population
  - does not represent mistakes, just the consequence of using samples instead of populations
- nonsampling error: result of poor sample design, sloppy data collection, faulty measuring instruments, bias, etc

EXPERIMENTS
experiments

- manipulating one or more variable by holding the others constant to find a relationship
- independent: variable being manipulated
  - has different values being tested
- dependent: responding variable
- experimental units: recipients of experimental treatment
  - could be anything
    ↳ people, plants, animals, etc.
- control: steps taken to reduce the effects of extraneous variables (other than independent / dependent)
  - control group: no treatment or neutral treatment
    ↳ used to compare
  - placebo: neutral treatment, no "real" effect aside from placebo effect
    ↳ sugar pill, etc
  - blinding: not telling who gets the placebo
    ↳ double blinding: experiment person also doesn't know
- randomization: randomly assigning experimental units to reduce extraneous variable effects on results
- replication: many experimental units to reduce variability
- confounding: occurs when the controls do not allow the experimenter to reasonably eliminate plausible explanations for an observed relationship
  - have controls

EXPERIMENTAL DESIGN
experimental design

1. identify the exact question and relevant population
2. plan for collecting representative data
   - observational study doesn't influence response
     ↳ confounding variables
   - experiment deliberately imposes something
3. collect data, minimize errors
4. analyze, draw conclusions, identify error sources

- design of an experiment describes the treatments and how the experimental units were assigned to the treatments
  - if experiments are poorly designed, we cannot see the effects of explanatory variables because they are confounded by other variables in the environment
    ↳ two factors are confounded when their effects on a response variable cannot be distinguished from each other
- randomized comparative experiment
  - similar groups of randomly assigned subjects before treatment
    ↳ no difference in experience so differences from the average response of groups must have been caused by the different treatments

- blocking is an experimental design component in which the researcher assumes that there are natural differences between categories within a block (gender, age, weight) and wants to eliminate the natural variation caused by this possible confounding variable
  - block design brings the people into the experiment as blocks and treatments are randomly assigned within the blocks
  - matched pair: experimental units are grouped into blocks of size two
- random sampling vs. random assignment
  - units within a block should look alike, units in one block should look different from units in another block
- lack of realism
  - difficult to realistically duplicate the actual conditions we want to study
  - subjects know they are in an experiment
- placebo effect: untreated subject believes they are receiving a real treatment and report an improvement
- hawthorne effect: treated subjects respond differently because they are a part of an experiment
- experimenter effect: researcher unintentionally influences subjects through factors such as facial expressions, tone of voice, or attitude

		allocation of units to treatment		
		random	not random	
selection of units	random	broad scope of inference, a random sample is selected and then randomly assigned	random sample is selected but there is no random assignment to treatments	A
	not random	narrow scope of inference. group is not selected at random but they are randomly assigned	observational study. no random selection or random assignment	B
		casual inference can be drawn	association, but no casual inference	

A: inference can be drawn to the population

B: inference is limited to the units included in the study

SAMPLING BIAS

sampling bias

- response bias: bias that results from problems in the measurement process
  - leading questions: loaded questions to favor one response over another
  - social desirability: responses may be biased to what one believes to be a desirable response
- increasing sample size cannot reduce survey bias, but it can reduce sampling error

OTHER BIASES

other biases

- availability bias: giving importance to memories most vivid and available for retrieval when unwarranted
  - distorts perception of past events and environment and complicates any attempts to make sense of it

sampling

- simple random sampling has a random sample of population
- systematic: every "x" person
- stratified: divide into groups based on similar characteristics and randomly select from each
- clustered: predetermined groups
  - cities, counties
- example: finding out soda preference outside of a school administration office
  - SRS might not work bc groups might have biases
  - systematic: choose every 6th person that comes by and ask if they prefer A or B
    ↳ alternate survey choices to prevent bias
  - draw conclusions based on findings

how would meskis inspect cars fairly?

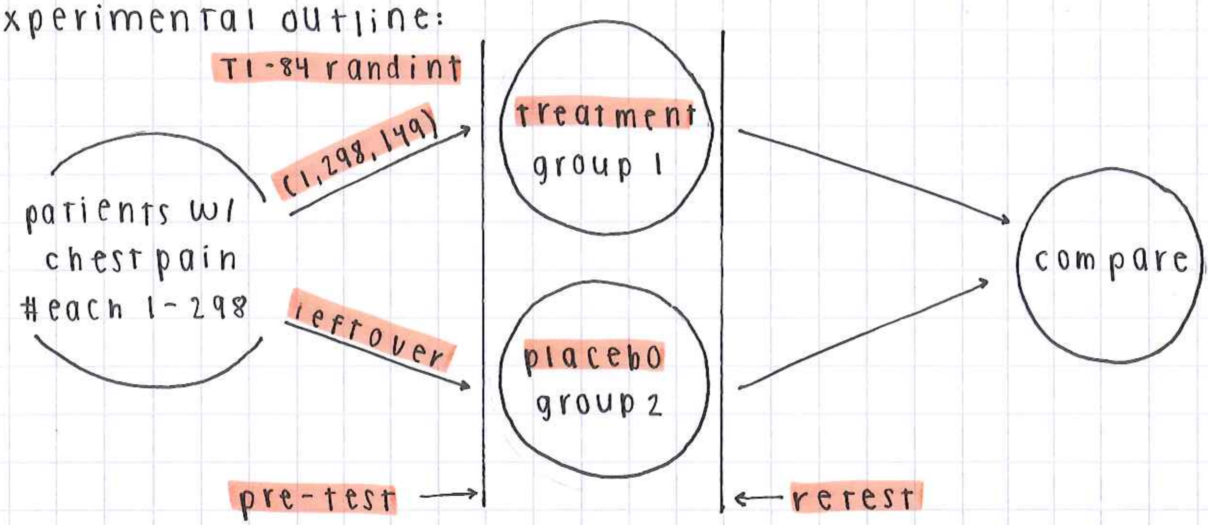
1. number each car from 1 to however many in the lot (ex. 1-500)
   - must be (001 - 500)
2. find a random spot on random # table
3. move in any direction by 3 digits until 7 are selected
   - skip #'s if not applicable (ex. 760)
4. inspect those cars

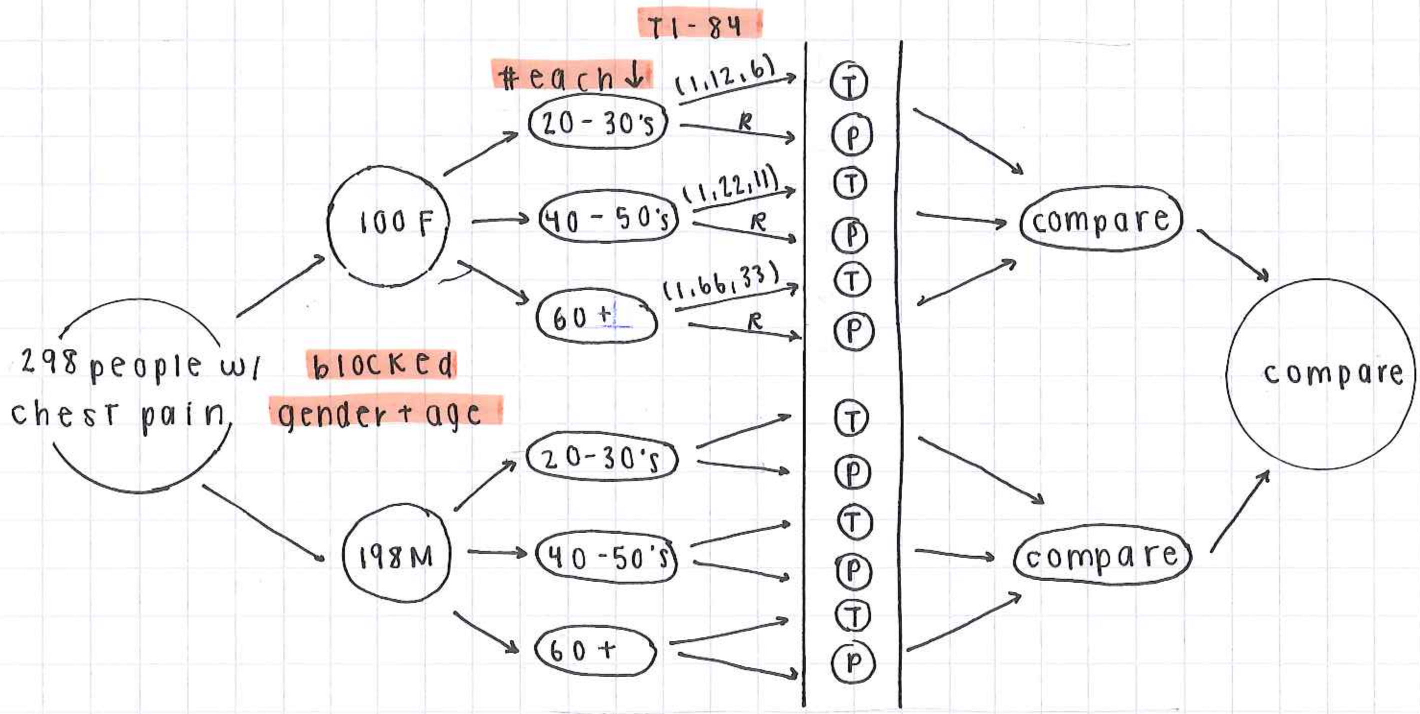
census + sampling

1. recent U.S. censuses are more accurate than early censuses
   - still not easy to get if it's optional or if they're out when the census is sent
     - ↳ minorities undercounted, college students elsewhere, homeless population underrepresented
2. censuses give people a voice, so an undercount would underrepresent the group. they lose money for social services because they don't get counted and given a voice

the experimental outline

experimental outline:





Emma Chau p.0

KEY TERMS

In an **observational study** researchers observe subjects and measure variables of interest. However, the researchers do not try to influence the responses. The purpose is to *describe* groups of subjects under different situations. In an **experimental study**, researchers deliberately apply some treatment to the subjects in order to observe their responses. The purpose is to study whether the treatment *causes* a change in the response.

In a **double-blind** experiment neither the subjects nor the individuals measuring the response know which subjects are assigned to which treatment. In a **single-blind** experiment the subjects do not know which treatment they are receiving but the individuals measuring the response do know which subjects were assigned to which treatments.

A **placebo** is something that is identical in appearance to the treatment received by the treatment group but has no effect.

A **control group** is an experimental group that does not receive the treatment under study. The control group could receive a placebo to hide the fact that no treatment is being given. In an **active control group**, the subjects receive what might be considered the existing standard treatment.

The explanatory variables in either an observational study or experiment are called **factors**. A **treatment** is any specific condition applied to the subjects in an experiment. If an experiment has more than one factor, then a treatment is a combination of specific values for each factor.

Two factors (explanatory variables) are **confounded** when their effects on a response variable are intertwined and cannot be distinguished from each other.

THE VIDEO

random assignment, treatment / control groups, large # of subjects, replication

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

manipulates subjects

1. Why is the study of the effect of humans on the coral reefs *not* an experiment?

it is an observational study because no treatment was imposed on the reefs. the researchers just observed the appearance of the reef

2. Who were the subjects in the Glucosamine/Chondroitin study? What did researchers want to find out?

subjects: patients with knee arthritis

reported reduction in knee pain = responding variable

3. Why were subjects randomly assigned to the treatments?

they cannot go into the experiment knowing what they received, and it made it so the groups were only as similar as chance could make them

4. Dr. Confound conducted a very badly designed experiment on mood-altering medication. List some of the problems with his experiment.

assessed in earshot of the other person

empathized with them

subject 7 gets to sit, not subject 8, for an hour

lack of blindness

leading questions

sample too small

laces day

- laces — no laces
  卌 卌 卌 卌 — 卌I
  卌 卌 卌 III
- pros:
  - got to see and observe real people
  - did stats
- cons:
  - not quality data collection
  - all male teachers
  - small sample size
  - not representative
    ↳ no list, no system for sampling
  - convenience sampling
  - rushed
  - people dress differently on different days
  - seasonal choices
  - wealth of students
  - absent / OPIS
  - trends
  - dress code
  - less ppl in zero period
- nonresponse bias vs. undercoverage bias
  ↳ done by responder ↳ fault of the statistician
  ↳ could be statistician's fault for not including enough in the survey
- response bias: recall or intentionally lying

- **Laces Done Well assignment ☆ not a project ☆**
  - make predictions; percentages and "more thans"
  - **population**: shoelace-wearing practices of ALL Oak Park high school students during early september of 2021
  - n = 80 + ; 20 freshman , 20 sophores, etc
    10 men, 10 women
  - use random selection or systematic
  - gather information evenly for the whole week and throughout the day
    ↳ early in the week, mid-week, end of week, weekend
  - don't ask; observation ONLY
  - write about difficulties
  - frequency tables, tallies, graphs
  - **conclusions**: why did we get the results?
  - discuss economic factors, location, ethnicity, age, time, weather, etc
    ↳ how these caused limitations on data
  - **reflection** on working with other people, process

simulations

1. describe the situation
   ex. coin flipping → 10 successive flips, getting 3 heads or tails in a row
2. state assumptions
   equal likelihood, independent of each other $P(H) = .5$, $P(T) = .5$
3. assign digits (random # table, calculator, computer) to represent the simulation
   evens = heads, odds = tails
4. simulate many times
   10 digits = 1 rep × 25 → one rep is NOT a simulation!
5. state your conclusions

example:

5.57)
1. basketball free throws, 5 free throws out of 5 shots, does she miss 3+?
   skill level = 70%
2. each shot is independent
   basket = .7, miss = .3
3. use RNT: 0-6 = make, 7-9 = miss
4. simulate looking for misses (using line 125)
5. 11/50, 20% of the time missed

simulations notes

mieck lecture

- simulations imitate real life situations
- 5.17 coin flip for babies
  - flip a coin until a head appears or until 4 flips have occurred, whichever comes first
  - enough repetitions might make it accurate

<u>Simulation</u>:

1. state the situation. what is the problem?
2. what are the assumptions?
3. assign digits to be similar to real outcomes
4. many repetitions (25+)
5. state conclusion; simulation is only approximate to the real thing

- coin flip simulation
  1. 3+ same (heads or tails)
  2. heads ≈ tails, equally likely → 50%, each coin toss is independent (not influenced by the other)
  3. RNT 0, 1, 2, ⋯ 9, independent
     1 digit = 1 coin flip → odd = H, even = T
     1, 3, 5, 7, 9 — 0, 2, 4, 6, 8
     $\boxed{\text{10 \#'s = 10 coin flips}}$ = 1 repetition
  4. repeat many times (25+)
  5. state conclusions

h	h	t	t	h	h	h				✓
1	9	2	2	3	9	5	0	3	4	

t	h	h	h							✓
0	5	7	5	6	2	8	7	1	3	

h	t	t	t							✓
9	6	4	0	9	1	2	5	3	1	

t	t	h	t	t	t	t	t	h	h	✓
4	2	5	4	4	8	2	8	5	3	

- 5.56 → abolishing exams
  1. state the problem: simulate surveying 10 univ. students about evening exams
  2. assumptions: Y/N are not equally likely → 80% favor abolishing get independent data (not in a group)
  3. assign digits: 0-9 are equal (1/10) and independent
     - 0 - 7 = in favor of abolishing exams
     - 8 + 9 = in favor of not abolishing
  4. 10 numbers at a time bc 10 university students look for all 10 saying yes.
  5. conclude

edpuzzle

- a couple plans on having 3 kids. design a simulation involving 20 trials that you can model the trials of the children

  use simulation to find probability of having 3 boys

  2/20 = 1/10 = .1 = 10%

  find the probability of having ONLY 2 girls

  6/20 = 3/10 = .3 = 30%

- there is a 60% chance of rain on Monday and 20% chance of rain on Tuesday

  - 1-6 = monday, 1-2 = tuesday

    any number that starts with 1-6 in the tens = monday, one's spot 1-2 = tuesday

surveys

good things

basic wording
even options for rating scale questions; never give a middle number option
reminder for people to log data if they keep the survey for 2+ days
ratio level data is best
choose current things to prevent recall bias
have clear, concise questions
help them feel anonymous and confidential (bag, no name or email listed, tell them)
help them give truthful answers (no peer influence, be neutral with reactions)
motivate them to respond (randomly selected, represents a portion of the population, incentivise)
eliminate bias

bad things

asking in front of a group
authority figure asking
verbal survey → should be written down
body language, facial expressions
suggestive, leading, seeking aggreement

simple "yes" and "no" don't show enthusiasm levels
"did you enjoy" "do you think"

not visually centered for rating scale questions, nor evenly spaced

AP STATISTICS: AGAINST ALL ODDS
Video 14 Worksheet

Name Emma Chau

SAMPLES AND SURVEYS

1. What is an estimate based on a sample? ______

2. What is a true value that describes an entire population? ______

3. What is the process of dividing a population into similar units? ______

4. What example of stratification is used in the video? ______

   How many strata are used? ______

5. In 1936, the Literary Digest predicted Alf Landon would win the presidential election. How many readers did the magazine poll? ______ How many people did Gallup poll? ______

   Who did Gallup predict as the winner? Roosevelt

   What was the problem with the magazine's poll? drew from a list that favored rich people

6. List three mistakes that can occur in polling.

   a. wording of a question → leading, loading info beforehand

   b. asking in a group setting

   c. using obscure vocabulary

   appearance matters

7. How many personal interviews are conducted each year as the core of the GSS? 1500

8. What is the histogram of the sampling process called? ______

9. What pattern does this distribution follow? ______

10. What is the peak of the distribution? ______

11. What happens to the distribution when the sample size is increased? ______

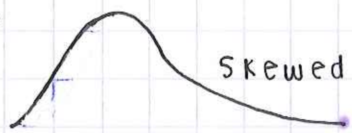
12. What determines precision? ______

basic graphs

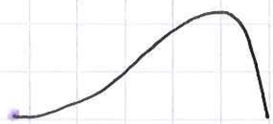
- median = 50th percentile
- graphs show info quickly, efficiently, attractively, but simply
  - start with a title
  - end with analysis; why did we get what we got?
- bar graphs
  - pareto chart (tall to small)
  - have labels on axes, scale bars properly; equal widths
- pictures should show proper area, volume, and shading
  - popcorn bucket might be 2x taller, but it's also 2x wider and deeper
    - ↳ ends up looking 8x bigger
- circle graphs / pie charts
  - comparing parts of a whole, percentages out of 100%.
  - use with categories
  - don't have distracting words everywhere
    - ↳ key, title, colors
- line graphs
  - tracking something over time
    - ↳ connect dots
  - double, triple, etc
  - squiggle at the bottom only
  - time plots track time over a period
- pictographs (symbols)
  - full symbol = amount in key, partial = less than
  - have equal size / spacing
- analysis: high, low, trend, why

dot plots, histograms

- florence nightingale (1820-1910)
  - one of the first nurses who used graphic reps of stats
  - reported on sanitary conditions; many of her recommendations were accepted
- dot plots
  - similar to bar graphs but with dots
  - good option for mid-sized data, not for data with high spread
  - title, label axis, equal spacing
  - centers: mean, median, mode
    ↳ identify which one used
  - shape: look for the tail
    ↳



skewed right



skewed left

  - outliers: deviant from center; look for gap
  - if you have plots side by side, analyze both
- histograms: bars touch
  - quantitative data grouped into classes
  - good for large amounts of data
  - problem: outliers get hidden in the categories
  - choosing how many bars / classes
    ↳ 5-10 usually; 7 is a good number
  - width of bars = $\frac{\text{max} - \text{min}}{\text{\# of classes}}$
    ↳ always round up
  - start with smallest data value, step up by width
  - use class limits, not boundaries
  - ex. step by 3, lowest # is 14
    ↳

+3	lower limit	upper limit	+3
	14	16	
	17	19	
	20	22	
	23	25	

  - frequency table: classes, tallies, frequency (# of tallies), relative frequency
  - mode is a good center
  - outliers: anything with relative freq. under 5%
- analysis: center, shape, spread, outlier

KEY TERMS

A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.

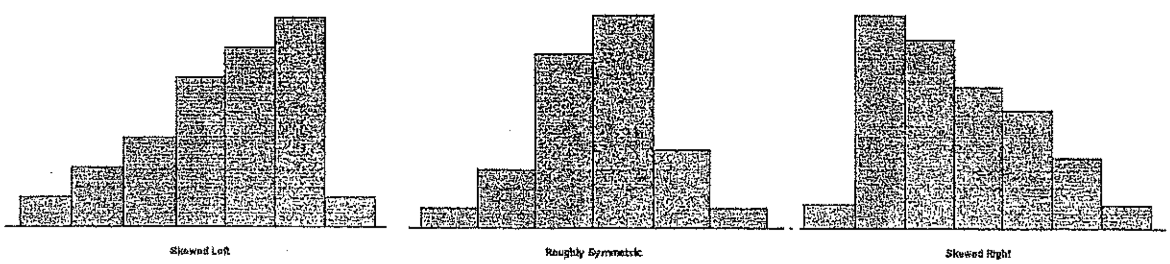


Figure 3.10. Shapes of histograms.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. The video opens by describing a study of lightning strikes in Colorado. What variable does the first histogram display?

first lightning strike times

2. In this lightning histogram, what does the horizontal scale represent? What does the vertical scale represent?

horizontal: time of day

vertical: % of days that the first flash occurred during that time

3. Was the overall shape of this histogram symmetric, skewed, or neither?

symmetric

4. Why were a few values in the second lightning histogram called outliers?

they stand out from the overall pattern of the histogram

5. When you choose the classes for a histogram, what property must the classes have if the histogram is to be correct?

there must be just enough to show data but not too much that information is unclear

6. What happens to a histogram if you use too many classes? What happens if you use too few?

too many: less informative because you can't focus in on data

too few: too generalized, not much information represented mean could be \$500,000 even if most houses sell for \$100 or \$200 thousand because of multimillion dollar homes

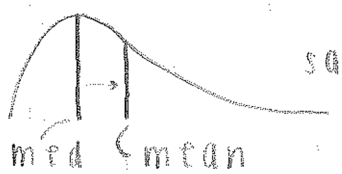
median could be \$160,000

4.2 percent

real income of top 1% grew by 17%. but this data showed that the growth eluded the lower, middle, and upper-middle class

symmetric: median = mean

unimodal with long upper tail:



same but opposite for lower tail

trimmed mean:

least to greatest, trim off trimming % of both sides

4.6: deleted 0, 0, 0 and 76, 123, 414   10%

10% trimmed mean $= \frac{0+0+\cdots+74}{24} = \frac{432}{24} = 18.0$

mean, $\bar{x} = 39.53$

4.1:

7.6, 8.3, 9.3, 9.4, 9.4, 9.7, 10.4, 11.5, 11.9, 15.2, 16.2, 20.4

mean =

median =

4.2: a) mean:

median

b) mean would ↓, median stay the same

c) trim % = 7.14%

stem + leaf plots

1. note max and min
2. draw a vertical line to split data values into stem | leaf
3. "let data flow" from stem to leaf
4. rake up the leaves (order from small to large. small near stem)
5. title, key, analysis

example: distances (in min) to get to school

- min = 3, max = 30

0 | 5 5 5 4 5 5 7 5 3 3 5 8 4 4 7 4 6
1 | 5 5
2 | 0 0 5 5 0 0
3 | 0 0 0

p.o distance to school (in min)

0 | 3 3 4 4 4 4 5 5 5 5 5 5 5 6 7 7 8
1 | 5 5
2 | 0 0 0 0 5 5
3 | 0 0 0

center: 5 (mode)
shape: skewed high
spread:
outlier: 30

Key: 2 | 5 = 25 min

- If back-to-back, do the following:

big ← small | 0 | small to big →
| 1 |
| 2 |
| 3 |

Key: 6 | 1 = 16
1 | 1 = 11

and do analysis for both!

C:
S:
S:
O:

measures of center

- measures of central tendency ; averages
  - mean, median, mode, trimmed mean
- mean (usually what people think of as "average"):
  - $\frac{\Sigma x}{n}$ $\frac{\text{(sum of numeric values)}}{\text{(\# of values)}}$
  - $\bar{x}$ for sample, μ for population
  - based on numerical values, so size matters
  - affected by outliers
- median: median value from small to large; "middle #"
  - position is emphasized, not numerical values
    - ↳ resistant to outliers
    - ↳ needs spread for context; not always a good representation
- mode: most often-occurring
  - categorical or quantitative
  - outliers have no effect
- trimmed mean: resists extremes
  - eliminates the pull of extremely low or high values (round up when calculating)
    - ↳ 10% trimmed mean of 20 values: remove lowest 2 and highest 2 numbers
    - ↳ data set should be in order first

calculator:

1. ordering a list
   [stat] → [edit] → [sort A]
2. finding mean, stdev, etc
   [stat] → [calc] → [1 var stats]

rounding rules:

0. _ _ _ _
1. _ _ _ _
2⁺. _ _

mean > median

- if mean is to the right of median = skewed right
- if mean is to the left, skewed left

mean < median

standard deviation

- measures of variation: how dispersed / spread / scattered the data is
  - range: high - low (DON'T USE - USELESS)
  - standard deviation: s (sample) or σ (population)
  - variance: s^2 (sample) or σ^2 (population)
  - coefficient of variation
  - interquartile range
  - chebychev's theorem
  - empirical rule

standard deviation:

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

squared bc otherwise sum = 0

ex.

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-4	16
4	-2	4
6	0	0
8	2	4
10	4	16

$\bar{x} = 6$

$n = 5$

$\Sigma = 40 / (n-1=4) = 10 = s^2$

↳ $\sqrt{s^2} = \sqrt{10}$, $s = 3.16$

- $\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$

- disadvantage of standard deviation: spread relies on unit of measurement, which makes it difficult to compare
  - CV = standard deviation of mean

    ↳ $CV = \frac{s}{\bar{x}} = \frac{\sigma}{\mu}$ ← stdev / ← mean

    ↳

0 - 10%	tight
13 - 33%	low
33 - 66%	medium
70 - 95%	high
100+ %	scattered

- chebychev's = anything bigger than 1 standard deviation
  - $1 - \frac{1}{K^2}$ ← # of standard deviations
  - at least 75% of data between $\mu - 2\sigma$ to $\mu + 2\sigma$
  - 88.9% between $\mu - 3\sigma$ to $\mu + 3\sigma$
  - 93.8% between $\mu - 4\sigma$ to $\mu + 4\sigma$

★ include sentence! ★

finding data spreads - 3.3

- range = smallest - largest
- variance and sample deviation: x = data value, n = sample size.
  > mean: average of data values
    ↳ $\bar{x} = \frac{\Sigma x}{n}$

    ↳ $x - \bar{x}$: difference between what happened and what you expected to happen; reps a "deviation" away
    ↳ $\Sigma (x-\bar{x})^2$: sum of squares; $\Sigma(x-\bar{x}) = 0$ bc negative cancels positive
  > sum of squares:
    ↳ defining formula: $\Sigma(x-\bar{x})^2$
    ↳ computation formula: $\Sigma x^2 - \frac{(\Sigma x)^2}{n}$

  > sample variance (s^2): average of $(x-\bar{x})^2$ values
    ↳ $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$ → estimate for population variance, usually no extreme values
    ↳ $s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}$
  > sample standard deviation: measure of variability or risk
    ↳ $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

    sq. rt to return to original units of data measurements
    ↳ $s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}}$
- population paramerers: N = data values of population
  > population mean $= \mu = \frac{\Sigma x}{N}$

  > population variance: $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$

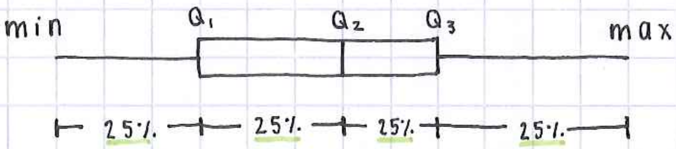
  > population standard deviation: $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$
- coefficient of variation (no units of measurement)
  > sample: $CV = \frac{s}{\bar{x}} \cdot 100\%$ > expresses standard deviation as a percentage

  > population: $CV = \frac{\sigma}{\mu} \cdot 100\%$

- chebyshev's theorem: for any set of data, population or sample, and for any constant k greater than 1, the proportion of the data that must lie within k st. dev. on either side is at least:
  > $1 - \frac{1}{k^2}$ at least 75% from $\mu - 2\sigma$ to $\mu + 2\sigma$
  88.9% from $\mu - 3\sigma$ to $\mu + 3\sigma$ 93.8% from $\mu - 4\sigma$ to $\mu + 4\sigma$

box and whisker

- mcdonald's fry assignment
  - sample
- 11, 16, 17, 19, 21, 21, 21, 21, 22, 23, 24, 24, 25, 25, 25, 25, 26, 26, 27, 27, 28, 28, 29, 30, 31, 35 ↙ ideal age to pause
  - med: 25
    - ↳ P_{50} (50th percentile): half of the data is above, half is below
- 95 percentile = only 5% got your result or higher
- quartiles split data set into 4
  - Q_1 = lower quartile (P_{25})
    Q_2 = P_{50} (median)
    Q_3 = upper quartile (P_{75})
- box and whisker: number line, title, axis labels, analysis



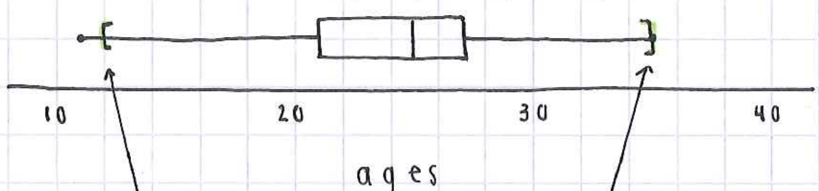
5 number summary (using #'s from above)

- min = 11
  max = 35
  Q_1 = 21
  Q_2 = 25
  Q_3 = 27

IQR = $Q_3 - Q_1$ = 27 − 21 = 6

C: 25 (med)
S: normal
S: 6 (IQR) "tight"
O: 11

the ideal age to pause



tukey and thresholds:
$Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$ are the "threshold limits"
$21 - 1.5(6) = 12$ $\quad$ $27 + 1.5(6) = 35$

calculator: $\boxed{\text{2nd}}$ $\boxed{\text{y=}}$ → choose plot → $\boxed{\text{zoom}}$ → $\boxed{\text{zoomstat}}$

KEY TERMS

A **five-number summary** of a set of data consists of the following:

minimum, first quartile (Q_1), median, third quartile (Q_3), maximum.

The **first quartile**, Q_1, is the one-quarter point in an ordered set of data. To compute Q_1, calculate the median of the lower half of the ordered data. The **third quartile**, Q_3, is the three-quarter point in an ordered set of data. To compute Q_3, calculate the median of the upper half of the ordered data.

A basic **boxplot** (or **box-and-whisker plot**) is a graphical representation of the five-number summary. A modified boxplot indicates outliers and adjusts the whiskers.

The **interquartile range** or **IQR** measures the spread of the middle half of the data:

$$IQR = Q_3 - Q_1$$

The **range** measures the spread of the data from its extremes:

range = maximum – minimum

THE VIDEO: BOXPLOTS

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What *variable* is used to compare different brands of hot dogs?

calorie content

2. What name do we give to the value for which one-quarter of the data values falls at or below it?

$Q_1 \rightarrow$ first quartile

3. What numbers make up a five-number summary?

minimum, Q_1, Q_2, Q_3, maximum

$Q_2 \downarrow$ median

4. How do you calculate the interquartile range?

$Q_3 - Q_1$

5. Boxplots show that poultry hot dogs as a group differ from all-beef hot dogs. Compare the distribution of calories between the two types of hot dogs.

75% of the poultry dogs have less calories than 75% of the beef dogs

at least half of the beef dogs have a calorie count higher than all of that of the poultry dogs

scatterplots

- line of best fit = least squares regression line
  - vertical distance from line = residual
    - ↳ residual = $y - y_p$
  - $\hat{y} = a + bx$
    - $a = \bar{y} - b\bar{x}$
    - $b = \dfrac{n\sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$
- extrapolation: below and above the min / max x-value, you can estimate using the line of best fit
  - usually isn't accurate past a few values
- when making the equation, make a T-chart with 2 points

x	y
$\bar{x}$	$\bar{y}$
0	a

OR

x	y
$\bar{x}$	$\bar{y}$
x	f(x)

✩ mark with something that's not a dot ✩

ex. 4.2 # 10 $\quad \hat{y} = a + bx, \quad a = \bar{y} - b\bar{x}$

x	y	x^2	xy
0	50	0	0
2	45	4	90
5	33	25	165
6	26	36	156

$\bar{x} = 3.25 \quad \bar{y} = 38.5 \quad \sum x^2 = 65 \quad \sum xy = 411$

$(\sum x)^2 = 169$

$\sum x = 13 \quad \sum y = 154$

$$b = \frac{4(411) - (13)(154)}{4(65) - (13)^2} = \frac{-358}{91} = -3.93$$

$a = \bar{y} - b\bar{x} = 38.5 - (-3.93)(3.25) = 51.27$

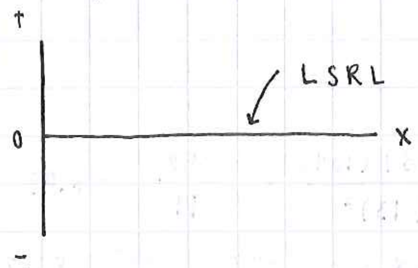
$\hat{y} = 51.27 - 3.93x$

x	y
3.25	38.5
0	51.27

- correlation coefficient: r ; $-1 \leq r \leq 1$ $\quad r \neq b$ $\quad r = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}}$
  - how close the dots "hug" the line
  - if $r = 1$, all dots are on the straight line ; perfect positive corr.
  - $r = -1$, perfect negative correlation
  - $r \approx 0$, no linear correlation
- rankings:
  - $.05 \leq r \leq .25$ "low correlation" / "weak"
  - $.3 \leq r \leq .65$ "medium"
  - $.7 \leq r \leq .85$ "high" / "strong"
  - $r \approx 1$ "very strong" / "near perfect"
- coefficient of determination: r^2 ; $0 \leq r^2 \leq 1$
  - if $r^2 = .9389$, that means:
    93.89% of the variation in the y (# of muggings) is explained by the least squares regression line and the variation in the x (# of uniformed police officers)
    6.1039% unexplained

- **outlier** has a large residual
- **influential observation** is a point close to the line that has a gap between it and the other data points
- **analysis:**
  Association (+/-)
  trend (as x increases, what does y do?)
  IN / OUT
  correlation (r-value and word)

- **residual plot:** if it shows a pattern, the original scatterplot is not linear
  - no pattern, x and y are linear



★ connect dots, label axes ★

Emma Chou p.0

KEY TERMS

Given a data set, one measure of center is the mean, $\bar{x}$. One way to judge the spread of the data is to look at the **deviations from the mean**, $x-\bar{x}$.

The **variance** is a measure of variability that is based on the square of the deviations from the mean. The formula for computing variance is:

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

Because the units for variance are the square of the units for the original data, we generally take the square root of the variance, which gives us the standard deviation:

$$s = \sqrt{s^2}$$

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. In comparing monthly precipitation for Portland, Oregon, and Montreal, Canada, why was comparing the mean monthly precipitation rates insufficient?

they were the same but weather patterns were different; one was more consistent throughout the year while the other had more concentrated seasons of rain

2. Why don't we measure spread about the mean by simply averaging $x-\bar{x}$, the deviations of individual data values from their mean?

some numbers are positive and some are negative, so by adding them all up to find the average, you'll get a sum of 0

3. What did the standard deviation of four-week sales data tell you about the two Wahoo's Taco locations, Manhattan Beach and South Coast Plaza?

Manhattan Beach had more deviation due to its dependence on the weather, while South Coast plaza had smaller deviation due to a steadier flow of customers at a mall

4. Can the standard deviation of a set of observations be $s=-1.5$? Explain.

no → standard deviation shows distance from the mean, so the distance can't be closer to the mean than 0 ins = point with a gap from other points but close to line

association, connection, correlation → rep. by variable r # and word

AP STATISTICS: AGAINST ALL ODDS
Video 8 Worksheet

Name Emma Chau P.6

outs = point with a large vertical gap

trend: as x ___, y ___

positive assoc. = direct

negative = indirect

DESCRIBING RELATIONSHIPS

1. What is a plot of quantitative variables? scatterplot

2. What is the x-variable called in studies? explanatory var. the y-variable? response var.

3. What is a variable that records into which of several categories a case falls? categorical

4. How do categorical variables enrich a scatterplot? add dimensions of information

5. What type of smoothing is found by slicing the scatterplot vertically, calculating the median within each slice, and connecting these medians by a straight line? median trace

6. What example in the video illustrates the use of a median trace? ______

7. What is the best fitting line that fits data by minimizing the sum of the squares of the residuals?

   ______

8. What example is used to illustrate the use of the least squares regression line? ______

9. In the equation y = a + bx, what is the formula for b? ______

   What is b in the equation? ______ What is the formula for a? ______

   What does y represent? ______ x? ______

   What is a in the equation? ______

10. Even though you can fit a regression line to any set of data, when is the line valid?

    ______

11. What are points with unusually large residuals? ______

12. What are points that deviate strongly in the x-direction? ______

AP STATISTICS: AGAINST ALL ODDS
Video 9 Worksheet

Name ______________________

CORRELATION

1. What is the measure of the strength and direction of the linear relationship between quantitative variables?

   coefficient of correlation → r

2. What values does r vary between? $-1 \leq r \leq 1$

3. What indicates a perfect positive correlation? 1 a perfect negative correlation? -1

4. What study in the video illustrates the use of correlation? how similar are traits of twins

   Which characteristics showed a strong correlation? height of twin A vs twin B

   Which characteristics showed a moderately strong correlation? personality of twin A vs. Twin B

5. In the formula for r, what do $\frac{x - \square}{s_x}$ and $\frac{y - y}{s_y}$ do? ______________________

   Why does the formula divide by n – 1? ______________________

   When is r positive? ______________________

   When is r negative? ______________________

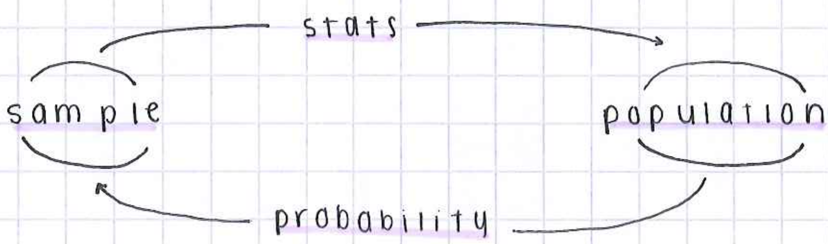
6. What kind of relationships does r measure? ______________________

7. What describes the amount of variation in y described by the linear relationship with x? ______________________

8. What example in the video uses the squared correlation coefficient? ______________________

probability

- **probability**: a measurement of the chance, or likelihood, of an event happening
  - $P(A)$ "p of A"
  - **probability** $= \frac{f}{n}$ $\frac{(\text{\# of desired})}{(\text{\# of total results})}$
  - $0 \leq p \leq 1$
    - ↳ $P(X) = 1$ "certain" $P(x) = 0$ "impossible"
- **complement of an event**: $P(\text{not } A) = P(A')$
  - $P(A) + P(A') = 1$; $P(A') = 1 - P(A)$



- **odds**: favorable to unfavorable
  - odds of rolling 6 on a die = 1:5
    - ↳ probability = 1/6
  - $F : F'$, $n = F + F'$
- **law of large numbers**: after many many trials, probabilities level out
- **sample space**: set of all possible outcomes
  - not # of outcomes

- **compound probability**: 2+ events occurring together
  - $P(A \cap B) = P(A) \times P(B)$
    - "and"
    - ↳ for independent events
  - $P(A \cap B) = P(A) \times P(B \mid A)$
    - "given that A occurred"
    - ↳ for dependent $\quad P(B) = P(B \mid A)$ if independent
- $P(A \cup B) = P(A) + P(B)$
  - "or"
  - for disjoint (mutually exclusive)
- $P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$
  - for non-disjoint (not mutually exclusive
- **mutually exclusive** = "no gaps, no overlaps"
  - probability sums to 1
  - "$\exists x$" = "there exists"

- combinations: ${}_nC_r$, order doesn't matter
  - ${}_nC_r = \dfrac{n!}{r!(n-r)!}$
    - ↳ ex. choosing a group of students
- permutations: ${}_nP_r$, order matters (much larger)
  - ${}_nP_r = \dfrac{n!}{(n-r)!}$
    - ↳ ex. first student gets x, second gets y, third gets z

~~you can no longer get divorced. And so the chance of that much bad luck is actually a little less than 1 in 250,000.~~

?→ Why multiply rather than add? Suppose you make a pack of trading cards out of the pictures of those 100 guys you've met so far through your Internet dating service, those men who in their Web site photos often look like Tom Cruise but in person more often resemble Danny DeVito. Suppose also that on the back of each card you list certain data about the men, such as honest (yes or no) and attractive (yes or no). Finally, suppose that 1 in 10 of the prospective soul mates rates a yes in each case. How many in your pack of 100 will pass the test on both counts? Let's take honest as the first trait (we could equally well have taken attractive). Since 1 in 10 cards lists a yes under honest, 10 of the 100 cards will qualify. Of those 10, how many are attractive? Again, 1 in 10, so now you are left with 1 card. The first 1 in 10 cuts the possibilities down by $\frac{1}{10}$, and so does the next 1 in 10, making the result 1 in 100. That's why you multiply. And if you have more requirements than just honest and attractive, you have to keep multiplying, so . . . well, good luck.

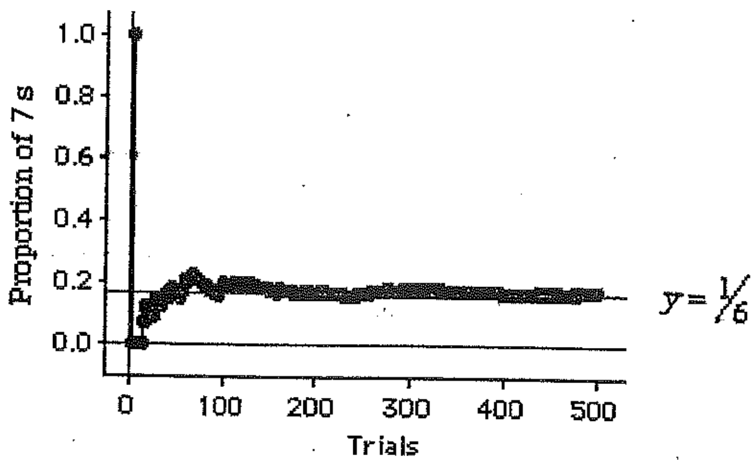
Before we move on, it is worth paying attention to an important detail: the clause that reads *if two possible events, A and B, are independent.* Suppose an airline has 1 seat left on a flight and 2 passengers have yet to show up. Suppose that from experience the airline knows there is a 2 in 3 chance a passenger who books a seat will arrive to claim it. Employing the multiplication rule, the gate attendant can conclude there is a $\frac{2}{3} \times \frac{2}{3}$ or about a 44 percent chance she will have to deal with an unhappy customer. The chance that neither customer will show and the plane will have to fly with an empty seat, on the other hand, is $\frac{1}{3} \times \frac{1}{3}$, or only about 11 percent. But that assumes the passengers are independent. If, say, they are traveling together, then the above analysis is wrong. The chances that both will show up are 2 in 3, the same as the chances that one will show up. It is important to remember that you get the compound probability from the simple ones by multiplying only if the events are in no way contingent on each other.

The rule we just applied could be applied to the Roman rule of half proofs: the chances of two independent half proofs' being wrong are 1 in 4, so two half proofs constitute three-fourths of a proof, not a whole proof. The Romans added where they should have multiplied.

There are situations in which probabilities *should* be added, and that is our next law. It arises when we want to know the chances of either one event or another occurring, as opposed to the earlier situation, in which we wanted to know the chance of one event and another event both happening. The law is this: *If an event can have a number of different and distinct possible outcomes, A, B, C, and so on, then the probability that either A or B will occur is equal to the sum of the individual probabilities of A and B, and the sum of the probabilities of all the possible outcomes (A, B, C, and so on) is 1 (that is, 100 percent).* When you want to know the chances that two independent events, A and B, will both occur, you multiply; if you want to know the chances that either of two mutually exclusive events, A or B, will occur, you add. Back to our airline: when should the gate attendant add the probabilities instead of multiplying them? Suppose she wants to know the chances that either both passengers or neither passenger will show up. In this case she should add the individual probabilities, which according to what we calculated above, would come to 55 percent.

These three laws, simple as they are, form much of the basis of probability theory. Properly applied, they can give us much insight into the workings of nature and the everyday world. We employ them in our everyday decision making all the time. But like the Roman lawmakers, we don't always use them correctly.

Example: In the casino game Craps, two dice are rolled and bets are made about the sum of the two dice. A bet that the next roll of the dice will show a sum of 7 pays 4:1 odds. Although the outcome on an individual roll of the dice is random, there is predictability in the long-run behavior. The graph shows the cumulative proportion of obtaining a 7 when two dice are rolled 500 times. Notice that the graph settles down and approaches the theoretical value of $\frac{1}{6}$.



- **The Law of Large Numbers:** The Law of Large Numbers states that the long-run relative frequency of repeated independent events gets closer and closer to the true relative frequency as the number of trials increases. We saw this law at work in the example above. In the long run, if you roll two dice many times, a sum of 7 will occur about $\frac{1}{6}$ of the time.
- **Independence:** Two events are independent if the occurrence of one event does not alter the probability that the other event occurs. If you roll two dice and obtain a sum of 7, the result of that roll has no effect on the next roll, so the two rolls are independent. But if you draw an ace from a deck of cards $P(\text{ace}) = \frac{4}{52}$ and without replacing it draw a second card, the probability that it is an ace is $P(\text{ace}) = \frac{3}{51}$. These events are *not* independent. (We will give a more formal definition of independence later.)

FORTUNE HUNTER

25 boxes, scratch 5 \$ icons

$P(\$ \text{ and } \$ \text{ and } \$ \text{ and } \$ \text{ and } \$)$
$= P(\$) \times P(\$ \mid \$) \times P(\$ \mid \$\$) \times P(\$ \mid \$\$\$) \times P(\$ \mid \$\$\$\$)$
$= \frac{5}{25} \times \frac{4}{24} \times \frac{3}{23} \times \frac{2}{22} \times \frac{1}{21} = \frac{120}{6375600} = .0000188$

given that one is chosen — 2 were chosen — 3 were chosen — 4 were chosen

to be almost certain, I need to do it 10^5 times → 100,000
→ 273.97 if one per day

BIRTHDAYS

7/27	12/21	12/14
2/21	7/14	12/23
2/29	7/2	9/16
1/21	12/1	10/15
12/7	6/1	
12/13	11/20	
10/15	3/15	

18 people, one pair of bdays
$2/18 = 1/9 = 11.11\%$

$\frac{1}{365} \times \frac{1}{364} \times \frac{1}{363} \times \frac{1}{362} \times \frac{1}{361} \times \frac{1}{360} \times \frac{1}{359} \times \frac{1}{358} \times$
$\frac{1}{357} \times \frac{1}{356} \times \frac{1}{355} \times \frac{1}{354} \times \frac{1}{353} \times \frac{1}{352} \times \frac{1}{351} \times \frac{1}{350} \times$
$\frac{1}{349} \times \frac{1}{348} = \frac{1}{8.}$

conditional prob.

contingency tables show conditional probabilities

- p(B) given p(A)

$P(B|A) = \frac{P(A \text{ and } B)}{P(B)}$ → $P(\text{Brwn Hair} | F) = \frac{P(\text{Brwn Hair} \cap F)}{P(F)} = \frac{3/10}{5/10} = \frac{3}{5}$ ← same

OR

out of 5 females $P(\text{Brwn Hair}) = 3/5$

$P(\text{survived} | \text{First}) = \frac{P(\text{survived} \cap \text{first})}{P(\text{First})} = \frac{203}{325} = 64.5\%$ medium chance of survival

$P(\text{survived} | \text{Third}) = \frac{P(\text{survived} \cap \text{third})}{P(\text{third})} = \frac{178}{706} = 25.2\%$ very unlikely chance to survive

$P(\text{crew} | \text{survived}) = \frac{P(\text{survived} \cap \text{crew})}{P(\text{survived})} = \frac{212}{711} = 29.8\%$ low chance of crew members survived

ex. medical tests: results can be positive or negative, whether or not the person has the disease (false positive or false negative)

	has condition	doesn't have	total
positive	110	20	130
negative	20	50	70
total	130	70	200

a) P(+ result | condition present)

$= \frac{P(+ \text{ result} \cap \text{condition present})}{P(\text{condition present})} = .8462 = 84.62\%$ "highly likely"

d) P(+ result | no condition) = $20/70$ = .2857 = 28.57%

e) P(condition present and positive) = $110/200$ = .55 = 55%

f) P(condition present and negative) = $20/200$ = .10 = 10%

a and b are conditionals, e and f are not!

poison ivy → c

P(N or M) are mutually exclusive because you can't have both a mild reaction and no reaction

- $P(A) + P(B) - \underbrace{P(A \text{ and } B)}_{0}$

independence is P(A and B) is P(A) = P(A|B)?

prob. distribution

- experiment: any process which a measurement is obtained
  - variable (x) ↳ quantitative
  - ex. amount of snowfall, weight of babies, etc
    not ex. nominal / ordinal data
- discrete random variables
  - finite, countable : -3, -2, -1, 0, 1, 2, 3, ...
  - no fractions, partial numbers, decimals
  - ex. # of students who voted, # of students who get A's, etc
- continuous random variables
  - infinite, countless : -3, -2, -1, 0, 1, 2, 3... but also partials
  - ex. temperature, height, inches of rainfall

discrete or continuous?
a) continuous
b) discrete
c) continuous
d) discrete

1) laugh
2) think
3) feel so deeply that you're moved to tears

– Jim Valvano

- probability distributions ↗ y-axis probability, x-axis variables
  - a graph which assigns probabilities to each random variables
  - histogram with discrete variables, density with continuous
  - all bars, all area, all probability + = 1
    ↳ all of the sample space, no gaps, no overlaps
    ↳ mutually exclusive
  - no class limits
    ↳ just discrete variable options

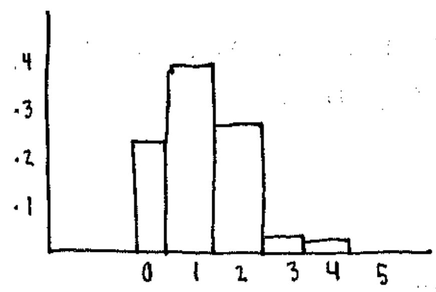
ex. boredom tolerance

- center and spread
  - distributions have expected values $\mu = \sum x \cdot p(x)$
  - standard deviation = σ = $SD(x) = \sqrt{\sum (x-\mu)^2 \, p(x)}$

6.1 #3. a) yes → probabilities add up to 1
b) no → probabilities all add up to 1.05 ; probably overlap

#12. d) $\mu = \sum x \cdot p(x) = 1.2530$
e) $\sigma = \sqrt{\sum (x-\mu)^2 P(x)} =$

put in var. and prob.
stat → calc → 1 var stats →


.4
.3
.2
.1
0
1
2
3
4
5

binom. distributions

1. jacob Bernoulli → swiss mathematician who studied Binomial experiments
2. binomial / Bernoulli experiments
3. probability of r successes out of n trials
4. F
5. n = 900
   >2 outcomes possible
   no → more than 2 outcomes possible bc he could get many different kinds of cars

success / Fail $p = r/n$

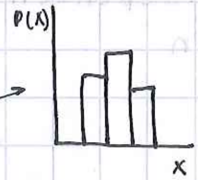
independent

discrete

-2, -1, 0, 1, 2 ...

continuous

-2, -1, 0 ; partials



$1 - p = q$

$\mu = \sum x \cdot P(X)$

$\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(X)}$

1. fixed # of trials, n
2. independent
3. 2 outcomes: success and fail
4. p = success, q = fail → $p + q = 1$ and $q = 1 - p$
5. find probability of r/n

binomial situations:

1. 72 trials
   independent
   S = 63, F = 9 success = < 3 min fail = > 3 min
   P(S) = .8 P(F) = .2 n = 72 r = 63+
2. trial = man being polled
   independent
   S = saying yes F = saying no
   P = .71, Q = .29
   n = 20, r = 18
3. not binomial; change the question

$$P(r) = \frac{n!}{(r!)(n-r)!} p^r q^{n-r} = C_{n,r}\, p^r q^{n-r}$$

6.2 hw (9, 10, 15, 16, 18-20, 27)

ex. n = 6, p = .70, q = .30, r = 4 success = germination, fail = no
$P(r) = C_{n,r}\, p^r q^{n-r} \rightarrow P(r=4) = C_{6,4}\,(.7)^4 (.3)^2 = .3241$
∃x 32.41% chance that exactly 4/6 tomato seeds will germinate

ex. eye operations

$P(r=5) = C_{6,5}(.3)^5(.7)^{6-5} = 6 \times .00243 \times .7 = .0102 \to 1.02$

∃x a 1.02% chance that 5/6 get their eyesight restored

- probability table: p, n

$P(r=6) \to n=6,\ r=6\quad p=.3\quad p(r=6)=.001$

- limitations: no other probabilities than increments of 5

- pop quiz problem: 10 Q's, a, b, c, d, e choices for each

$P(r \geq 8 \text{ correct}) = ?$ assume independence

- trial = answering each question

S = right, F = wrong

$p = 1/5 = .2\quad Q = .8$

$n = 10\quad r = 8, 9, 10$

$P(8) = 0\quad P(9) = 0\quad P(10) = 0$

$P(r=8) + P(r=9) + P(r=10) = .000 + .000 + .000 = 0$

∃x a 0% that a student would blindly guess 8+ correctly on the pop quiz

6.2 #18 $p = .1 \to 10\%$

a) $P(r=0)$

2nd vars → binompdf

trials: 7

p: .10

x value (r): 0

binompdf(7, .1, 0) = .4783 ← DON'T WRITE THAT ON AP EXAM

$P(r=0) = .4783$

b) $P(\text{at least } 1) \to P(r \geq 1) = 1 - P(r=0) = 1 - .4783 = .5217$

c) $P(\text{no more than } 2) \to P(r \leq 2) = P(r=0) + P(r=1) + P(r=2)$

$= .4783 + .3720 + .1240 = .9743$

2nd vars → binomcdf ← cumulative

X value: 2 binomcdf (7, .1, 2) = .9743

6.2 #20

a) $p = .10\quad n = 20$

$P(r \leq 1) = 1 - P(r=0) = .1216$

$P(r \leq 1) = .8784 \to$ ∃x an 87.84% chance that at least 1...

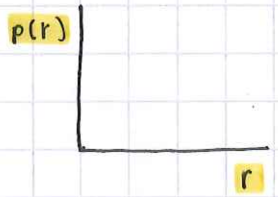
b) $P(r > 2)$ binomcdf (20, .1, 2) = .6769

$P(r > 2) = 1 - .6769 = .3231 \to$ ∃x a 32.31% chance that more than 2...

c) $P(r=0) = .1216 \to$ ∃x a 12.16% chance that no ties...

d) $P(\text{at least 18 not too tight}) = P(r \leq 2) \to$ binomcdf(20, .1, .2) = .6769

binomial distribution

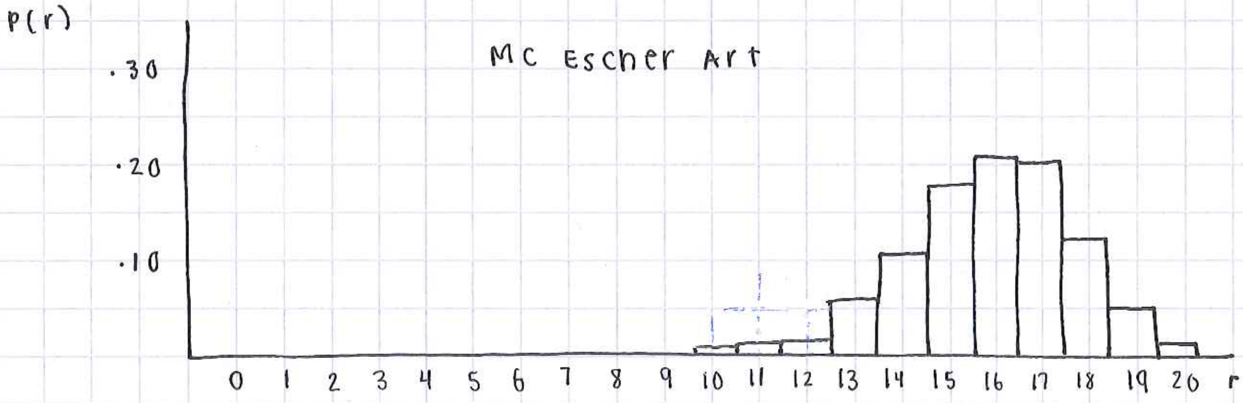


$n=6, \quad p=.25$

- $\mu = n \cdot p$
  $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{\mu q}$

art appreciation class: pass or fail
$p = .8$
$n = 20$
assume independence

- $P(r=10) = .002$
- $P(r=0) = .000$



C: 16 (mode) $= \mu$ $\quad \sigma = 1.7889$
S: skewed low
S: $CV = \sigma/\mu = .1118 = 11.18\%$ "right"
O: 0 – 11, 20

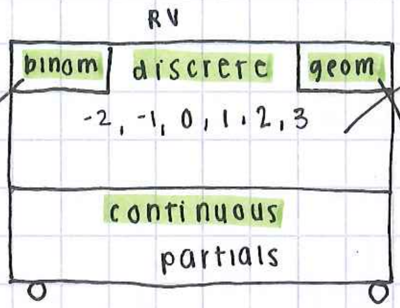
6. a) .3784
.1791
.2466
.1892
.0068

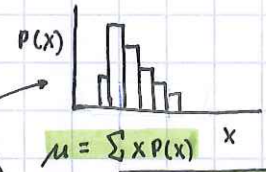
b) $\mu = \sum x\, P(x)$ $\quad \mu = 5.28$ $\quad \sigma = 4.88$ $\quad \sigma/\mu = .9242$ "high"

geom. distributions

S = make, F = miss   p = .83, q = .17   r = 4, n = 5

geometric distributions
- going **until** 1 success; n is unknown and $n \neq 0$

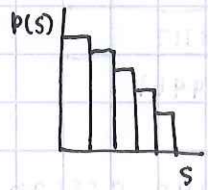




$\mu = \sum x P(x)$
$\sigma = \sqrt{\sum (x-\mu)^2 P(x)}$

$\mu = np$
$\sigma = \sqrt{(1-p)\mu} = \sqrt{\mu q}$

S/F, going **until** 1st success



always skewed right
X is never 0
$\mu = 1/p$
$\sigma = \sqrt{q}/p$

geometric settings:
1. 2 outcomes → S/F
2. p = prob. of success
3. independent
4. variable of interest = # of trials until success → no fixed amt

$P(X=n) = (1-p)^{n-1}p = q^{n-1}p$

x	1	2	3	4	5	...
p(X)	p	(1-p)p	q^2p	q^3p	q^4p	...

← infinite

graph is a downward stair step b/c you're always multiplying by something less than 1

$P(X>n) = (1-p)^n = q^n$
- $P(X>12) = q^{12}$

8.26
a) independent   p = .03   q = .97
S = defective hard disk drive, F = works
b) $P(X=5)$

x	1	2	3	4	5
P(X)	.03	.0291	.0282	.0274	.0266
		$q \cdot p$	q^2p	q^3p	q^4p

$P(X=5) = .0266$
∃x a 2.66% chance that a hard disk drive will be found defective on trial 5

$P(X=5) = q^{n-1}p = (.97)^4(.03) = .0266$

8.41

a)

h	h	h
h	h	t
h	t	h
t	h	h
t	t	h
t	h	t
h	t	t
t	t	t

$p = 2/8 = 1/4 = .25$

b) $p = 6/8 = 3/4 = .75$

c) x = # of trials until a winner

d)

x	1	2	3	4
P(x)	.75	.1875	.0469	.0117
cdf	.75	.9375	.9844	.9961

collegeboard vid ✲ always assign random variable ✲

H = # of tropical storms until first hurricaine, p = .41

$P(H=4) = (1-.41)^3(.41) = .795$ $\mu = 1/p$ $\sigma = \sqrt{q}/p$

$P(x=4) =$ geompdf $(p=.53, x=4) = .0842 \rightarrow$ must include a sentence!

.2 of the litters have 4+ pups

$P(x=5) = q^{n-1}p = (.8)^4(.2) = .0819 \rightarrow \exists x$ an 8.19% chance that the fifth litter will be the first to have 4+ pups

$\mu = 1/p = 1/.2 = 5$ litters

powerade problem: ind. x = trials until winning $p = .25, q = .75$

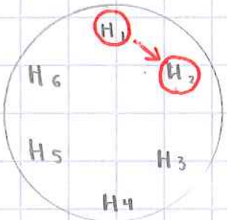
\# of trials?

calculator: geomet pdf (.25, 17) = .0025 → ∃x a .25% chance that you win on the 17th time

hermit problem

emma chau p.2

infected

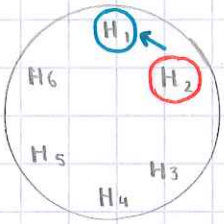


probabilities change; not binomial / geom

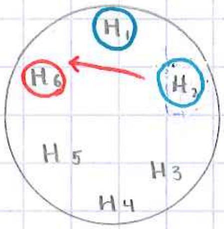
$P(0) = 0$, $P(1) = 0$

$P(2) = 1/5 \rightarrow$ can visit any of the 5

immune



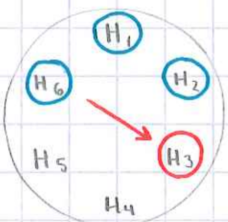
if H_2 goes back to H_1 only 2 were infected



H_2 infects non-immune (4/5 chance)

if H_6 randomly visits an immune hermit (2/5 chance), only 3 will be infected

$P(3 \text{ infected}) = 4/5 \cdot 2/5 = 8/25$

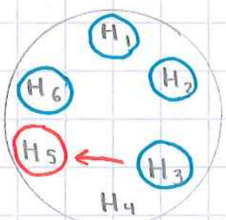


H_6 randomly infects H_3 (non-immune) (3/5 chance)

if H_3 visits an immune hermit (3/5 chance)

$P(H_2$ infects non-immune$) \cdot (H_6$ infects non-immune$) \cdot (H_3$ visits immune$)$

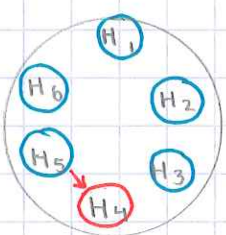
$P(4) = (4/5) \cdot (3/5) \cdot (3/5) = 36/125$



H_3 infects non-immune (2/5 chance)

if H_5 visits immune (4/5 chance), disease ends

$P(5) = 4/5 \cdot 3/5 \cdot (H_3$ infects non-immune$) \cdot$
$(H_3$ visits immune$) = 4/5 \cdot 3/5 \cdot 2/5 \cdot 4/5 = 96/625$



H_5 infects non-immune (1/5 chance)

H_4 visits immune (5/5 chance)

$P(\text{all infected}) = 4/5 \cdot 3/5 \cdot 2/5 \cdot 1/5 \cdot 5/5 = 24/625$

$\mu = \Sigma x p x = 0 + 0 + 2(1/5) + 3(8/25) + 4(36/125) + 5(96/625) + 6(24/625)$
$= 2194/625 = 3.51$ hermits