# Lecture 01

## Instructor Information:

## Lecture/Workshop Information:

## Webpage Information:

## Textbook Information:

# General Outline of Course:

# Marking Scheme:

# Some Hints:

# Lecture 02

**Deterministic vs Stochastic Systems:**

- **classical laws of physics → deterministic**

- **a coin flip → stochastic?**

- **why don't some systems repeat themselves?**

- **stochastic systems are often convenient**

## Examples of Statistical Practice:

- **sample surveys - results of opinion polls**

- **business - selling airline tickets?**

- **agriculture - how to optimize yield?**

- **population biology - how many fish?**

- **education - comparing learning techniques**

- **sports - handicapping in golf**

- **sports - when should the goalie be pulled?**

- **health - longitudinal studies**

- **experimental design**

## 1. Descriptive Statistics:

- **addresses the following problem**
  - **given some data, try to understand it**

- **the data can be a *sample* or a *population***
  - **eg: the weights of STAT270 students in kg**

- **descriptive statistics is summarization**

- **summaries can be *numerical* or *graphical***
  - **eg:**

## 2. Inferential Statistics:

- **addresses the following problem**
  - **given a sample, try to understand popln**

- **mathematical vs inferential reasoning**
  - **mathematical reasoning (general → specific)**
  - **inferential reasoning (specific → general)**
  - **eg**
  - **eg**

- **inferential reasoning uses probability theory**

# Lecture 03

## Dotplots:

- a graphical descriptive statistic
- applicable given univariate data $x_1, \ldots, x_n$
- able to observe centrality, dispersion, outliers
- not so widely used (histograms are better)

**Histograms:**

- **a graphical descriptive statistic**
- **applicable given univariate data** $x_1, \ldots, x_n$
- **able to observe centrality, dispersion, outliers**
- **we encourage intervals of equal length**
- **generated by statistical software**

**Histograms (we illustrate by hand):**

- **data are weights of students in kg: 47, 55, 79, 63, 64, 67, 54, 59, 58, 84, 70, 61, 65, 59**

## Issues in constructing histograms:

- always label axes and provide a title
- how many intervals should be chosen?
- be aware of the scale of the vertical axis
- handling intervals that are not of equal length

**Sample mean $\bar{x}$:**

- **a numerical descriptive statistic of centrality**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x_i}{n}$

**Sample median $\tilde{x}$:**

- **a numerical descriptive statistic of centrality**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \textbf{if } n \textbf{ odd} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right)/2 & \textbf{if } n \textbf{ even} \end{cases}$

**Consider a sample of $n$ house prices:**

- $\bar{x} = \$850,000$
- $\tilde{x} = \$700,000$
- **Why do the statistics differ?**

**The median is more *robust* than the mean wrt outliers:**

**Know how to approximate the median and mean from a histogram:**

# Lecture 04

**Variability (dispersion) in data:**

- **Consider the following two datasets**
  - **Dataset 1: -2, -1, 0, 1, 2**
  - **Dataset 2: -300, -100, 0, 100, 300**

**Sample range $R$:**

- **a numerical descriptive statistic of variability**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $R = x_{(n)} - x_{(1)}$
- **not so widely used anymore**
- **based on only two data values**

**Sample variance $s^2$:**

- **a numerical descriptive statistic of variability**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad = \quad \frac{(\sum x_i^2) - n\bar{x}^2}{n-1}$
- $s^2 \geq 0$; $\qquad s^2 = 0$ **corresponds to** $x_1 = \cdots = x_n$
- **large $s^2$ corresponds to widely spread data**
- **note that denominator is $n - 1$ instead of $n$**
- **think about why the difference $x_i - x_n$ is squared**
- **distinguish between the two formulae**
- **note that $s^2$ is measured in squared units**
- **the sample standard deviation is given by $s$**

## How do location/scale changes affect $\bar{x}$ and $s^2$:

- **i.e.** $x_i \to y_i = a + bx_i$

- **e.g. changing Celsius data to Fahrenheit**

**Problem: Can you construct a dataset with $R = 30$ and $s^2 = 100$?**

**Problem:** $n = 5$**,** $x_1 = 10, x_2 = 3, x_3 = 7, x_4 = 8$

**(a) If** $\bar{x} = 6$**, obtain** $x_5$**.**

**(b) If** $s = 5$**, obtain** $x_5$**.**

## Boxplots:

- a graphical descriptive statistic
- applicable given univariate data (in groups)
- generated by statistical software
- calculations require $\tilde{x}$, lower fourth, upper fourth
- interpreting boxplots is our focus
- boxplots are not as popular as they should be

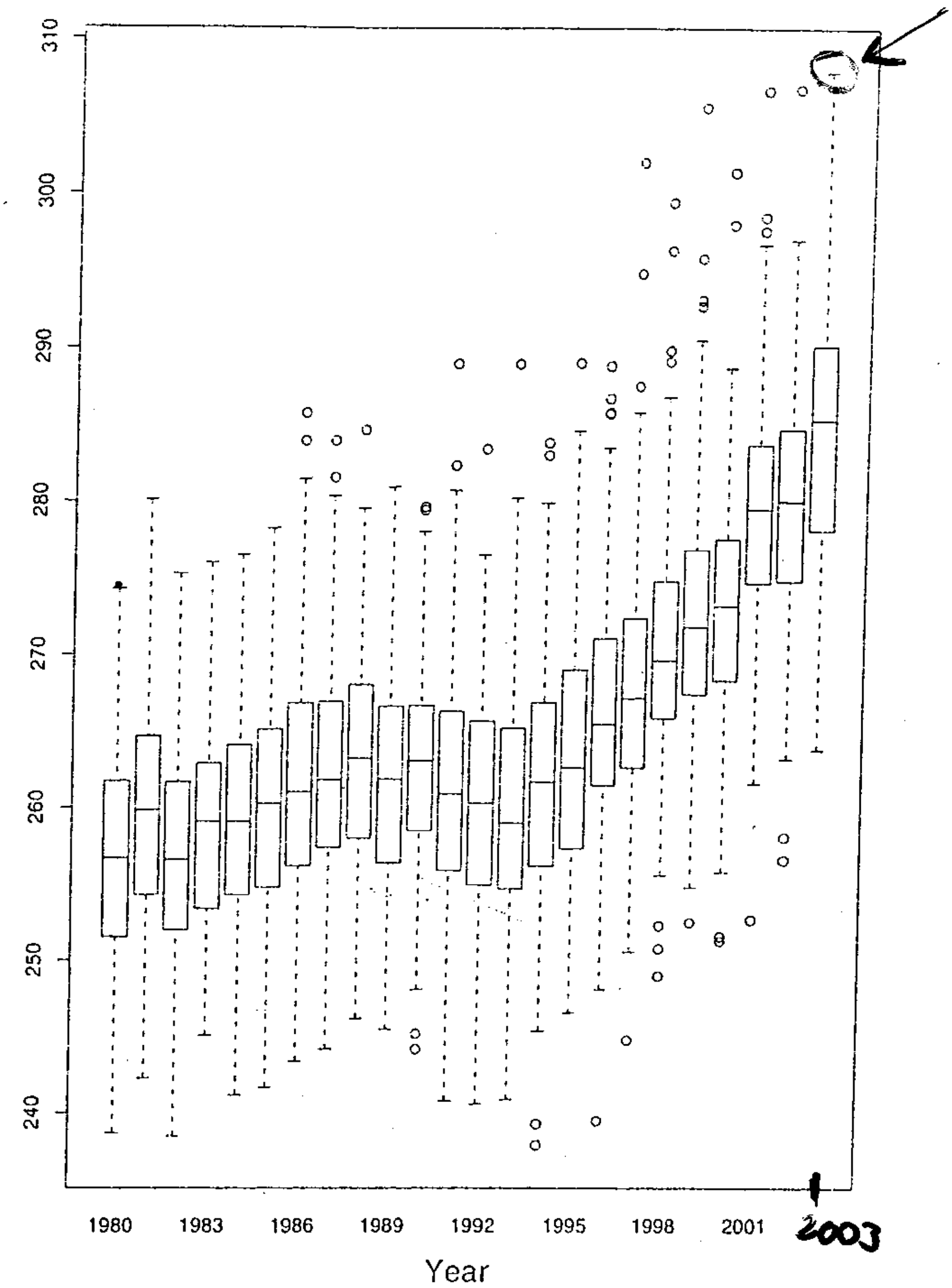


**Figure 1. Boxplots for driving distances from 1980 through 2003.**

# Lecture 05

## Scatterplots:

- **a graphical descriptive statistic**

- **for paired quantitative data** $(x_1, y_1), \ldots, (x_n, y_n)$

- **always label axes and provide a title**

- **focus is on the relationship between** $x$ **and** $y$

- **scatterplots aid in prediction**

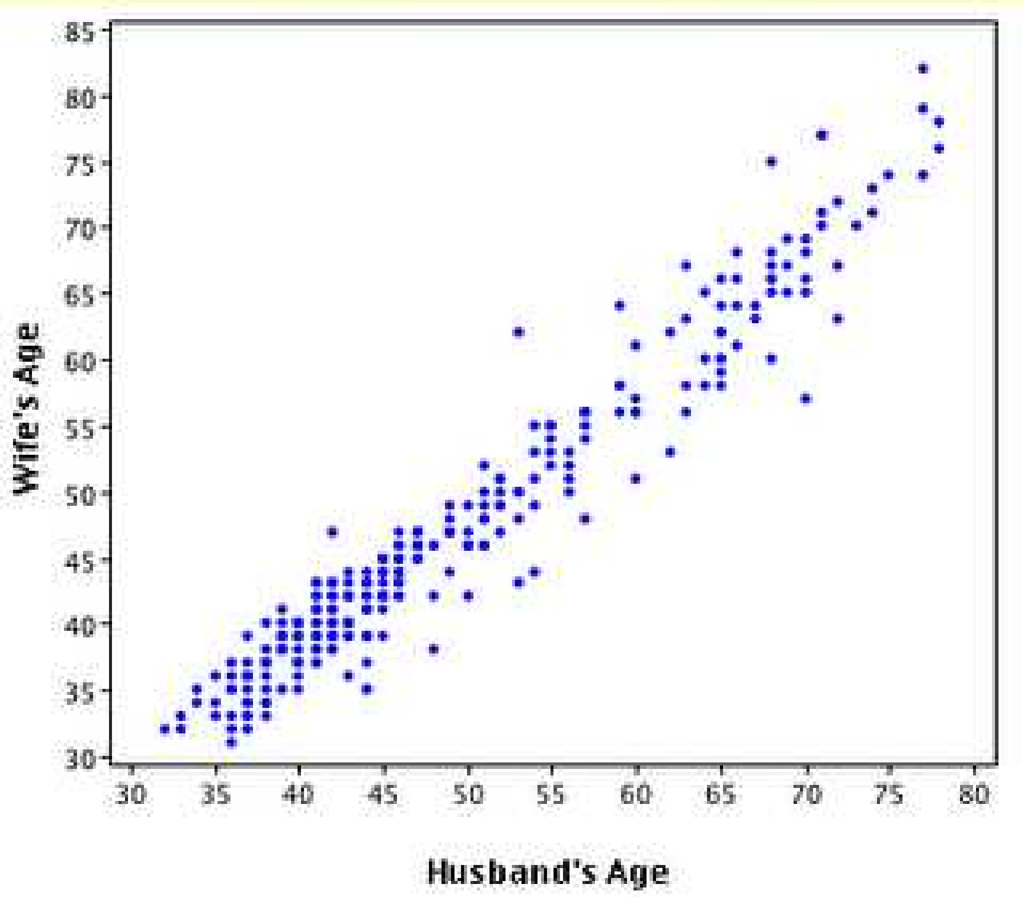
- **interpolation versus extrapolation**

**Examples: data appropriate for a scatterplot?**

**(a) Consider 20 patients who take drug 1 and we record their blood pressure ($x$'s). There are 20 other patients who take drug 2 and we record their blood pressure ($y$'s).**

**(b) Consider the monthly immigration rates ($x$'s) into British Columbia and the monthly emigration rates from British Columbia ($y$'s).**

**(c) We consider 10 different colours. In a neighbourhood, we count the number of houses of each colour.**


Wife's Age
85
80
75
70
65
60
55
50
45
40
35
30
30
35
40
45
50
55
60
65
70
75
80
Husband's Age

**Sample correlation coefficient $r$:**

- **a numerical descriptive statistic**
- **for paired quantitative data** $(x_1, y_1) \ldots, (x_n, y_n)$
- $r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$
- $r$ **describes linearity between** $x$ **and** $y$

**Association versus cause-effect:**

- **correlation does not imply causation**
- **the role of lurking variables in causation**
- **observational studies**
- **randomized experiments**

**Example for discussion: "Prayer can Lower Blood Pressure", USA Today, August 11, 1998.**

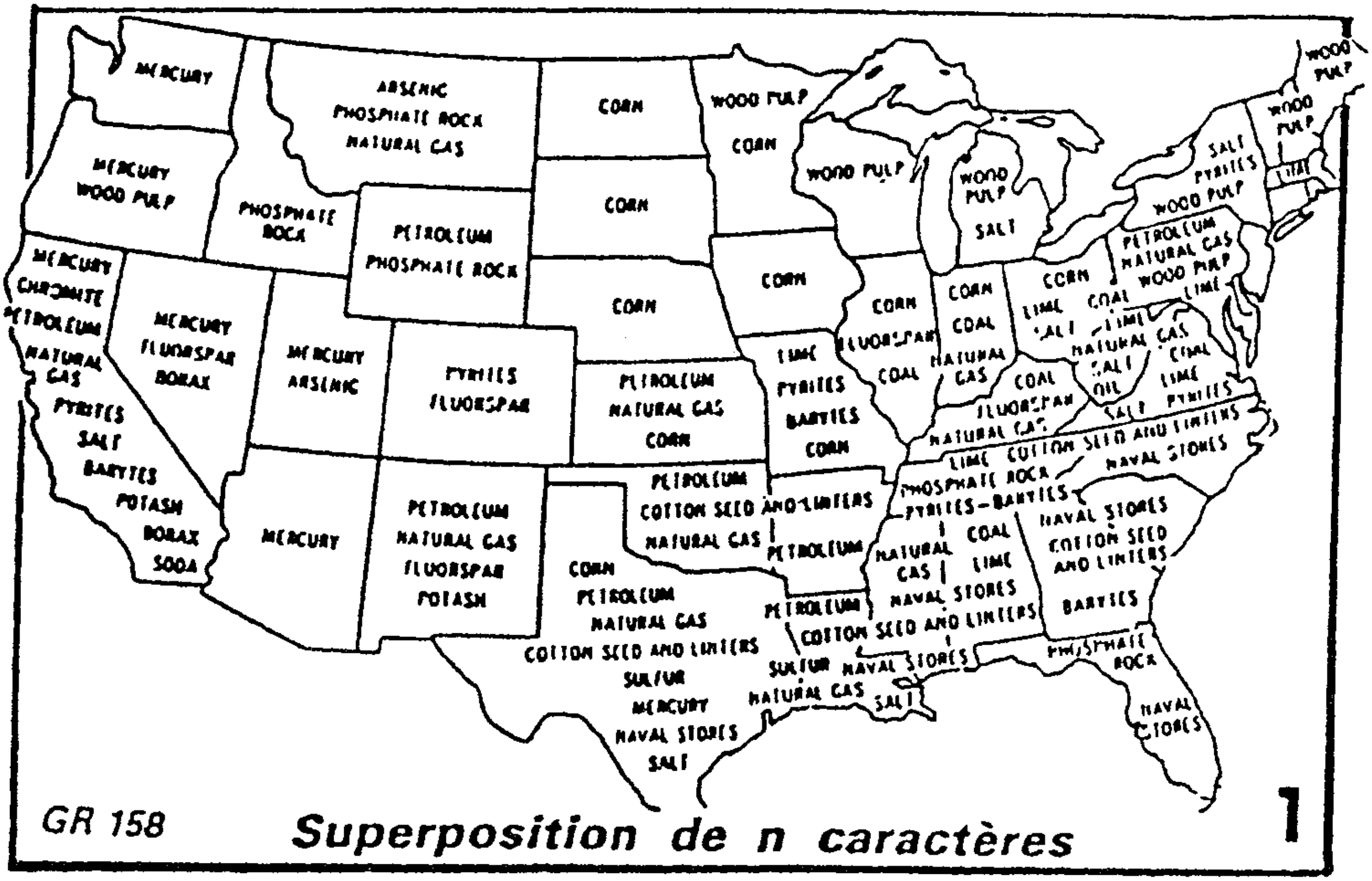
**People who attended a religious service once a week and prayed or studied the Bible were 40% less likely to have high blood pressure.**

# Lecture 06

**More on graphical statistics:**
**Recall that the purpose of a graphical descriptive statistic is to facilitate insight with respect to the dataset. Although there are various standard graphical statistics (e.g. histograms, boxplots, scatterplots), sometimes data with a non-standard structure may benefit from a special-purpose graphical display.**

**The only limit in developing graphical displays is your imagination. Keep in mind however that the goal is to learn from the display. Therefore, simplicity and clarity are important considerations. On the following pages, we give an example of a non-standard dataset and special purpose graphical displays that aid in addressing various questions.**

# Raw Materials for Chemical Industry



Question 1: Which states produce particular raw materials?

Question 2: Which raw materials do states produce?

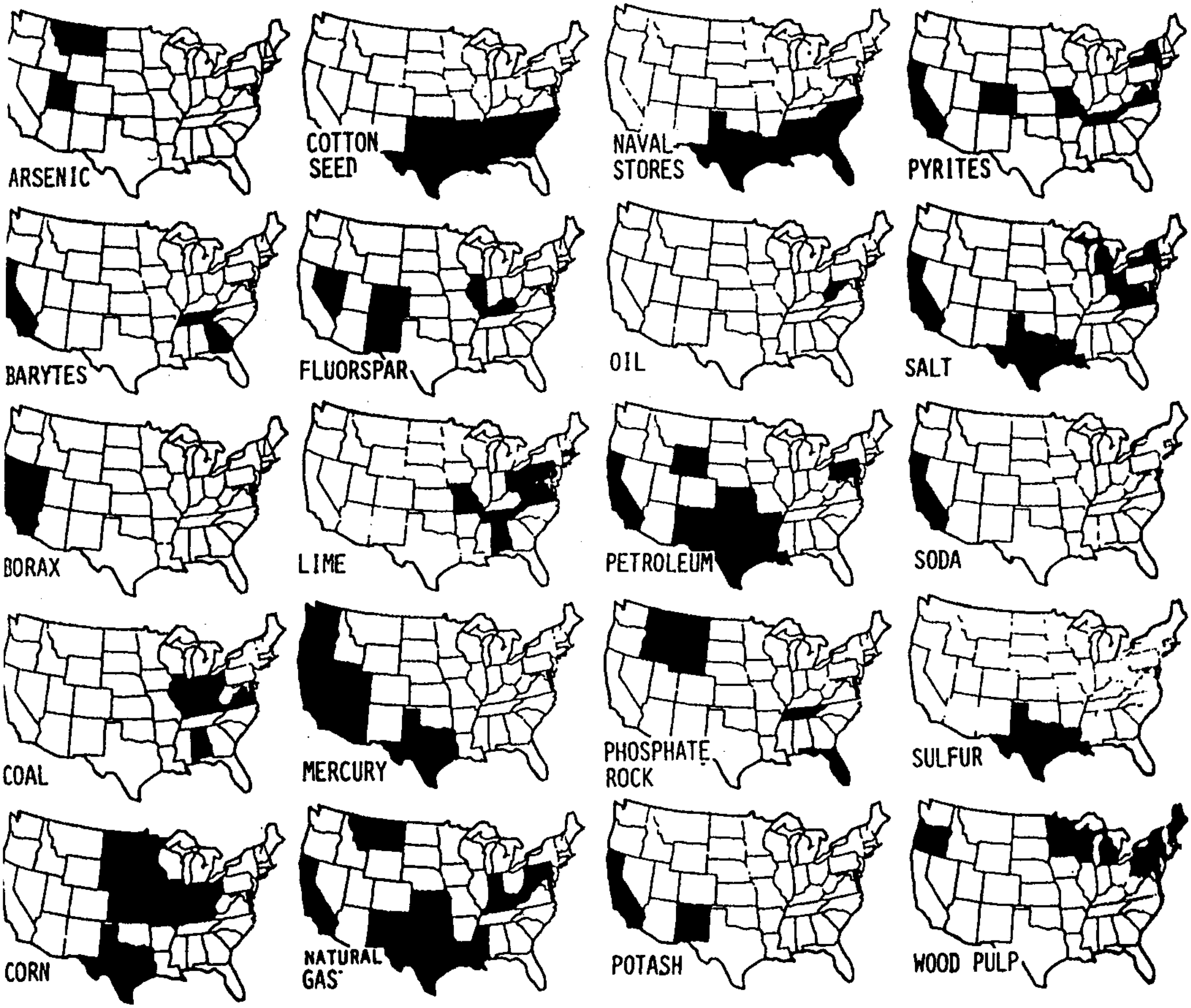


Figure 3. A collection of the one-variable maps arranged alphabetically.

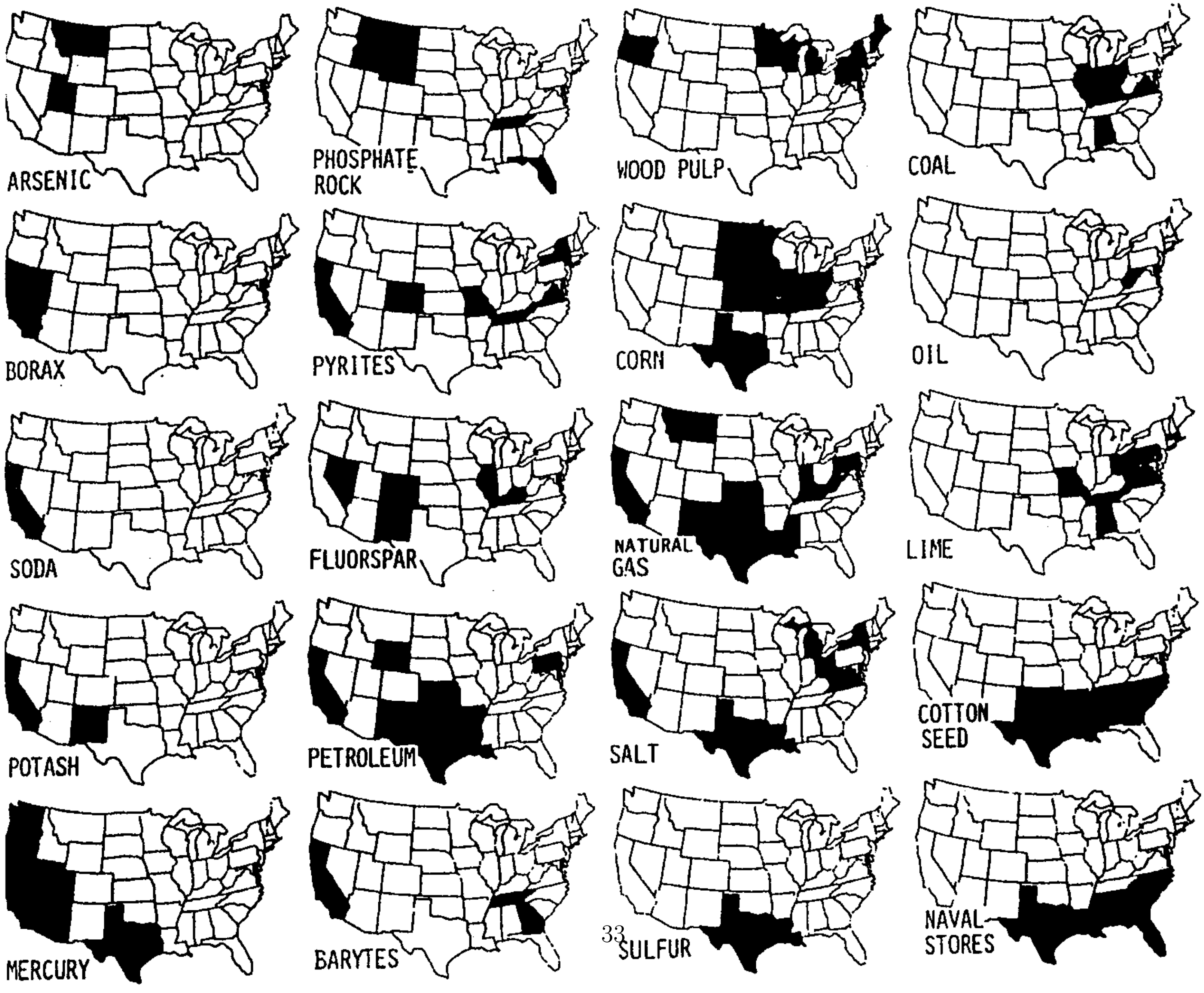


**Figure 4. A collection of the one-variable maps arranged geographically.**

**Introduction to probability:**
**We think of an *experiment* as any action that produces data.**

**The *sample space* is the set of all possible outcomes of the experiment.**

**An *event* is a subset of the sample space.**

**Example 1: flipping a coin three times.**

## Example 2: total auto accidents in BC in a year

## Example 3: lifespan in hours of 2 components

**Problem: Write down the sample space wrt the experiment where you roll a die until an even number occurs.**

**Set theory for events using Venn diagrams:**

- **"$A$ union $B$"** $\equiv$ "$A$ **or** $B$" $\equiv A \cup B$

- **"$A$ intersect $B$"** $\equiv$ "$A$ **and** $B$"$\equiv A \cap B \equiv AB$

- **"$A$ complement"** $\equiv \bar{A} \equiv A' \equiv A^c$

**Definition:** $A$ **and** $B$ **are *mutually exclusive* (*disjoint*) if** $A \cap B = \phi$**.**

**DeMorgan's Law:** $\overline{A \cup B} = \bar{A} \cap \bar{B}$

**Something to think about: Probability is used in everyday language yet it is not well defined. What is meant by the statement "the probability of rain today is 0.7"?**

**Oxford English Dictionary definition of probability: extent to which an event is likely to occur, measured by the ratio of favourable cases to the whole number of cases possible**

**Kolmogorov (1933) provided the following definition of probability:**

**A probability measure $P$ satisfies three axioms**

1. **For any event $A$, $P(A) \geq 0$**
2. **$P(S) = 1$ where $S$ is the sample space**
3. **If $A_1, A_2, \ldots,$ are disjoint, $P(\cup A_i) = \Sigma P(A_i)$**

**Discussion points:**

## Useful derivations from the Kolmogorov defn:

**Example:** $P(\bar{A}) = 1 - P(A)$

**Example:** $P(\phi) = 0$

**Example: If** $A \subseteq B$**,** $P(A) \leq P(B)$

**Example:** $P(A \cup B) = P(A) + P(B) - P(AB)$

**Example:**

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & -P(AB) - P(AC) - P(BC) \\ & +P(ABC) \end{aligned}$$

# Lecture 07

**Symmetry definition of probability:**

**In the case of a finite number of equally likely outcomes in an experiment,**

$$P(A) = \frac{\textbf{number of outcomes leading to } A}{\textbf{number of outcomes in the experiment}}$$

**Example: Roll two dice. Let $A$ be the event that the sum is 10.**

**Discussion points:**

**Frequency definition of probability:**

**In hypothetical identical trials of an experiment,**

$P(A) =$ **the long term relative frequency of** $A$

**Example: Roll two dice** $n$ **times. Let** $A$ **be the event that the sum is 10.**

**Discussion points:**

**Problem: If 85% of Canadians like either baseball or hockey, 63% like hockey and 52% like baseball, what is the probability that a randomly chosen Canadian likes both hockey and baseball?**

**Conditional probability (an important topic):**

**The conditional probability of $A$ given $B$ is**

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

**provided that $P(B) \neq 0$.**

**Problem: Suppose that I roll a die and tell you that the result is even. What is the probability that the outcome is a 6?**

**Problem: The probability of surviving a transplant operation is 0.55. If a patient survives the operation, the probability that the body rejects the transplant within a month is 0.2. What is the probability of surviving both critical stages?**

**Confusion of the inverse:** $P(A \mid B) \neq P(B \mid A)$

**A patient has a lump in her breast. A physician believes that there is a 1% chance that the lump is malignant. A mammogram is positive where mammograms are accurate 80% of the time when lumps are malignant and mammograms are accurate 90% of the time when lumps are benign. The test comes back positive.**

**What is your opinion concerning the probability of the malignancy of the lump?**

**Problem: In each box of my favourite cereal, there is a prize. Suppose that the cereal company distributes 10 different prizes randomly in the boxes of cereal. If I purchase five boxes of cereal, what is the probability that I obtain five different prizes?**

**The Monty Hall problem: On the game show "Let Make a Deal", a contestant is given the choice of three doors. Behind one door is a grand prize (e.g. a car) and behind the other two doors are gag gifts. The contestant picks a door, and Monty (who knows what is behind all of the doors), reveals a gag gift by opening one of the two doors that the contestant has not chosen. Monty then gives the contestant the choice of switching doors between the remaining two unopened doors. Should the contestant switch?**

# Lecture 08

**Independence: Lets begin thinking about independence in an informal way. Two events are independent if the occurrence or nonoccurrence of one event does not affect the probability of the other event.**

**Formally, and this is how you are required to prove independence, events $A$ and $B$ are independent if and only if**

$$P(AB) = P(A)P(B)$$

**Example: Suppose that I flip a coin and roll a die. What is the probability of obtaining a tail and a six?**

**Topic for discussion: Suppose that you go to a casino and you are watching roulette. You are thinking about placing a bet on either red or black. You have observed that the roulette wheel has resulted in a black number 6 times in a row. Do you bet red or black?**

**Does your opinion change if black comes up 100 times in a row?**

**More on independence: There is a connection between conditional probability and independence.**

**Proposition: Suppose $P(A) \neq 0$, $P(B) \neq 0$ and $A$ and $B$ are independent. Then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.**

**The converse is also true.**

**Definition: Events $A_1, \dots, A_k$ are *mutually independent* if and only if the probability of the intersection of any $2, 3, \dots, k$ of these events equals the product of their respective probabilities.**

**Example: Consider the case of mutual independence of the events $A_1, A_2, A_3$ and $A_4$.**

**Example of pairwise independence but not mutual independence: Roll two dice and define**

- $A_1 \equiv$ **first die is odd**
- $A_2 \equiv$ **second die is odd**
- $A_3 \equiv$ **sum of both dice is odd**

**The birthday problem: Amongst 30 people, what is the probability that at least two of them share a common birthday?**

**Generalize the problem to $n$ people.**

**Basic combinatorial results:**

**Proposition: The number of *permutations* of $n$ distinct objects is $n! = n(n-1)(n-2)\cdots 1$**

**Example: We can permute symbols A, B and C in $3! = 6$ ways.**

**Definition:** $0! = 1$**.**

**Proposition: The number of permutations of $r$ objects chosen from $n$ distinct objects is $n^{(r)} = n!/(n-r)!$**

**Example: We can permute two of the symbols A, B, C, D and E in $5^{(2)} = 5!/(5-2)! = 120/6 = 20$ ways.**

# Lecture 09

**Proposition: The number of *combinations* of $r$ objects chosen from $n$ distinct objects is**

$$\binom{n}{r} = \frac{n^{(r)}}{r!} = \frac{n!}{r!(n-r)!}$$

**Example: We can choose two of the symbols A, B, C, D and E in** $\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$ **ways.**

**Calculating combinations by hand: Try $\begin{pmatrix} 30 \\ 4 \end{pmatrix}$.**

**Example: There are 20 people in a room. How many committees of four people can be chosen?**

**Anecdote regarding combination locks:**

**Proposition: There are $\binom{n}{r}$ ways of partitioning $n$ distinct objects into a first group of size $r$ and a second group of size $n-r$.**

**Corollary:** $\binom{n}{r} = \binom{n}{n-r}$

**Proposition: Let $n = n_1 + \cdots + n_k$. There are $\frac{n!}{n_1!n_2!\cdots n_k!}$ ways of partitioning $n$ distinct objects into $k$ distinct groups of sizes $n_1, n_2, \ldots, n_k$.**

**Example: How many ways can we partition the symbols A, B, C and D into distinct groups of sizes 1, 2 and 1?**

**Lets summarize: We have been developing counting rules, specifically $n!$, $n^{(r)}$, $\left(\begin{array}{c} n \\ r \end{array}\right)$ and $\frac{n!}{n_1!n_2!\cdots n_k!}$.**

**Whereas none of these rules are too difficult individually, the challenge is to use the counting rules to calculate probabilities.**

**Using the symmetry definition, recall that the probability of an event $A$ is given by**

$$P(A) = \frac{\textbf{number of outcomes where } A \textbf{ occurs}}{\textbf{total outcomes in the experiment}}$$

**Problem: In a class of 100 students, 20 are female. If we randomly draw five students to form a committee, what is the probability that at least two of the committee members are female?**

**Problem: In a row of four seats, two couples randomly sit down. What is the probability that nobody sits beside their partner?**

**Problem: We roll a die. If we obtain a 6, we choose a ball from box A where three balls are white and two are black. If we do not obtain a 6, we choose a ball from box B where two balls are white and four are black.**

**(a) What is the probability of obtaining a white ball?**

**(b) If a white ball is chosen, what is the probability that it came from box A?**

**Problem: Five cards are dealt from a deck of 52 playing cards. What is the probability of**

**(a) three of a kind?**

**(b) two pair?**

**(c) straight flush?**

# Lecture 10

## More probability calculations:

## Lotto 649:

- $P($**jackpot**$)$
- $P($**five matching numbers**$)$
- $P($**four matching numbers**$)$
- $P($**two matching numbers**$)$

# Keno calculations: similar to Lotto 649

**Problem: There are $N$ people who attend the theatre and check their coats. At the end of the performance, the coats are randomly returned. What is the probability that nobody receives their own coat?**

**Problem: Out of 300 woodpeckers, 30 have damage to the beak but not the crest, 50 have damage to the crest but not the beak and 10 have damage to both the beak and crest.**

**(a) How many woodpeckers have no damage?**

**(b) For a randomly chosen woodpecker, are crest and beak damage independent?**

**(c) For a randomly chosen woodpecker, are crest and beak damage mutually exclusive?**

**Problem: Consider 12 balls (3 orange, 3 green, 3 blue and 3 red). We randomly choose 9 balls from the 12. How many different looking selections can be made?**

**Problem: How many bridge hands are there?**

**Problem: If we scramble the letters R, O, T, T, N, O, and O, what is the probability that we spell TORONTO?**

**Problem: A batch of 100 stereos contains $n$ defective speakers. A sample of 5 stereos is inspected. What is the probability that $y$ stereos are defective?**

**Problem: Consider a bag with 4 red marbles and 6 black marbles. What is the probability of obtaining 3 red marbles if we draw three marbles from the bag (i) with replacement and (ii) without replacement? Repeat the calculations if the bag contains 40 red and 60 black marbles.**

# Lecture 11

**Coincidences are often misunderstood.**

**Example for discussion: Richard Baker left a shopping mall, found what he thought was his car and drove away. Later, he realized it was the wrong car and returned to the parking lot. The car belonged to another Mr Baker who had the same model of car, with an identical key! Police estimated the odds of this happening at one million to one.**

- **Were the police correct?**
- **How astonished should we be?**

**Example for discussion: Consider the case of twins who were separated at birth. They later meet as adults and are amazed that they share some striking characteristics (eg. they use the same toothpaste, their eldest children have the same names, they have the same job).**

**Should they be amazed?**

**Definition: A *random variable* (rv) is a function of the sample space.**

**Example: A coin is flipped three times. Let $X$ be the number of heads.**

**Definition: A random variable is *discrete* if its outcomes are discrete.**

**Definition: A random variable that takes on the values 0 and 1 is *Bernoulli*.**

**Example: Consider the temperature in degrees Celsius. Let $Y = 1(0)$ if the temperature is freezing (not freezing).**

**Definition: The *probability mass function* (pmf) of a discrete random variable $X$ is**

$$p_X(X = x) = P(s \in S \ : X(s) = x)$$

**Example: Consider the experiment consisting of three flips of a coin. Let $X \equiv$ the number of heads. Obtain the pmf for $X$.**

**Proposition: A pmf $p_X$ satisfies**

**(1)** $p_X(x) \geq 0$

**(2)** $\Sigma_x \, p_X(x) = 1$

**Example: Let $X$ be the sum of two dice. Obtain the pmf of $X$.**

**Example: Consider a batter with a 300 average. Let $X$ be the number of at bats until the batter gets a hit. Obtain the pmf of $X$.**

**Problem: A library subscribes to two weekly magazines, each of which is suppose to arrive on Wednesdays. In actuality, the two magazines arrive independently with probabilities of arrival,** $P(\text{Wed}) = 0.3$**,** $P(\text{Thu}) = 0.4$**,** $P(\text{Fri}) = 0.2$ **and** $P(\text{Sat}) = 0.1$**. Let** $Y$ **be the number of days beyond Wednesday that it takes for both magazines to arrive. Obtain the pmf of** $Y$**.**

**Problem: At the end of an exam, four textbooks are left behind. At the beginning of the next lecture, the four texts are randomly returned to the four students. Let $X$ be the number of students who receive their own book. Obtain the pmf of $X$.**

# Lecture 12

**Definition: The *cumulative distribution function* (cdf) of a random variable $X$ with probability measure $P_X$ is given by**

$$F_X(x) = P_X(X \leq x)$$

**Example: Consider three flips of a coin and let $X$ be the number of heads. Obtain the cdf of $X$.**

**Properties of a cdf $F$:**

**(1) $F$ is normed (i.e. $F(-\infty) = 0$, $F(\infty) = 1$)**

**(2) $F$ is monotone increasing**

**(3) $F$ is right continuous**

**Given a cdf corresponding to a discrete distribution, be able to determine the pmf.**

**Example:**

**Definition: The *expectation* of a discrete rv $X$ with pmf $p(x)$ is given by**

$$\mu \equiv \mathrm{E}(X) \equiv \sum_x x \ p(x)$$

**The expectation can be thought of the long run average of the random variable over hypothetical repetitions of the experiment.**

**Example: Consider the experiment consisting of three flips of a coin. Let $X \equiv$ the number of heads. Obtain $\mathrm{E}(X)$.**

**Example: Consider the experiment of tossing a die and let $X$ be the outcome. Obtain $\mathrm{E}(X)$.**

**Earlier it was stated that we view the expectation of a random $X$ as the long run average of $X$. Lets explore this statement by considering $N$ hypothetical repetitions of the experiment.**

**Proposition: The expectation of a function $g(X)$ corresponding to the discrete random variable $X$ with pmf $p(x)$ is given by**

$$\mathrm{E}(g(X)) = \sum_x \; g(x) \; p(x)$$

**Example: Consider the experiment of tossing a die and let $X$ be the outcome. Obtain $\mathrm{E}(X^2)$.**

**Proposition:** $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$

**Problem: A store orders copies of a weekly magazine for its magazine rack. Let $X$ be the weekly demand for the magazine with pmf**

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $p(x)$ | $\frac{1}{15}$ | $\frac{2}{15}$ | $\frac{3}{15}$ | $\frac{4}{15}$ | $\frac{3}{15}$ | $\frac{2}{15}$ |

**Suppose that the store owner pays \$1 for each copy of the magazine and the customer price is \$2. If leftover magazines at the end of the week have no salvage value, is it better for the owner to order three magazines or four magazines?**

## Is expectation always a reasonable criterion?

**Problem for discussion: Suppose that you are given the chance to play a game a single time where the entrance fee is $1 million dollars. With probability 0.99, you lose and receive nothing. With probability 0.01, you win and receive $1 billion dollars. Should you play the game?**

# Lecture 13

**We explore *expectation* in more detail.**

**Proposition: For a discrete rv $X$ with pmf $p(x)$**

$$\mathrm{E}(g_1(x) + \cdots + g_k(x)) = \mathrm{E}(g_1(x)) + \cdots + \mathrm{E}(g_k(x))$$

**Definition: The *variance* of a discrete rv $X$ with pmf $p(x)$ is**

$$\sigma^2 \equiv \mathrm{V}(X) \equiv E[(X - \mathrm{E}(X))^2]$$

- **we call $\sigma$ the *standard deviation***
- **$\sigma$ and $\sigma^2$ are measures of spread**
- **contrast sample quantities $(\bar{x}, s)$ with popln quantities $(\mu, \sigma)$**

**Example: Consider the experiment consisting of three flips of a coin. Let $X \equiv$ the number of heads. Obtain $\mathrm{V}(X)$.**

**Proposition:** $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$

**Proposition:** $\mathrm{V}(aX + b) = a^2\mathrm{V}(X)$

**Example: Let $X$ be the average january temperature in degrees Celsius where $\mathrm{E}(X) = 5$C and $\mathrm{V}(X) = 3(\mathrm{C})^2$. Find the expected value and the variance of $Y$ where $Y$ is the average january temperature in degrees Fahrenheit.**

**Problem: Calculate $\sigma$ and $\mathrm{E}(3X + 4X^2)$ corresponding to the rv $X$ with pmf $p(x)$ where**

| $x$ | 4 | 8 | 10 |
|---|---|---|---|
| $p(x)$ | 0.2 | 0.7 | 0.1 |

**Example: In a game of chance, I bet $x$ dollars. With probability $p$, I win $y$ dollars. What should $x$ be for this to be a fair game?**

**Definition: A discrete rv $X$ has a *Binomial* distribution denoted $\mathrm{Bin}(n, \theta)$ if it has pmf**

$$p(x) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

**Motivation for the Binomial - the most important discrete distribution:**

**Consider performing an experiment $n$ times where the probability of success in every trial is $\theta$ and the $n$ experiments are independent. We are interested in the probability of $x$ successes.**

**The probability of getting $x$ successes $S$ and $n-x$ failures $F$ in the specific order**

$$SS\cdots S \quad FF\cdots F$$

**is**

$$\begin{aligned} & \theta\theta\cdots\theta \quad (1-\theta)(1-\theta)\cdots(1-\theta) \\ = \ & \theta^x(1-\theta)^{n-x}. \end{aligned}$$

**Therefore**

$$P(x \text{ successes}) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

**for $x = 0, 1, \ldots, n$.**

**Key points for the Binomial distribution:**

1. **the $n$ trials are independent**
2. **same probability of success $\theta$ in each trial**

**Example: You roll a die 10 times and are interested in obtaining 5's or 6's. What is the probability that $x$ rolls result in either 5's or 6's?**

**For** $X \sim \mathrm{Bin}(n, \theta)$

- $\mathrm{E}(X) = n\theta$
- $\mathrm{V}(X) = n\theta(1-\theta)$

## Proof (first result):

$$\begin{aligned}
\mathrm{E}(X) &= \Sigma_{x=0}^{n}\ x \left(\begin{array}{c} n \\ x \end{array}\right) \theta^{x}(1-\theta)^{n-x} \\
&= \Sigma_{x=1}^{n}\ x \tfrac{n!}{x!(n-x)!}\ \theta^{x}(1-\theta)^{n-x} \\
&= n\theta\ \Sigma_{x=1}^{n} \tfrac{(n-1)!}{(x-1)!(n-x)!}\ \theta^{x-1}(1-\theta)^{n-x} \\
&= n\theta\ \Sigma_{y=0}^{n-1} \tfrac{(n-1)!}{y!(n-1-y)!}\ \theta^{y}(1-\theta)^{n-1-y} \\
&= n\theta\ \Sigma_{y=0}^{n-1} \left(\begin{array}{c} n-1 \\ y \end{array}\right) \theta^{y}(1-\theta)^{n-1-y} \\
&= n\theta
\end{aligned}$$

# Lecture 14

**Problem: A friend recently planned a camping trip. He had two flashlights, one that required a single 6-V battery and another that used two size-D batteries. He had previously packed two 6-V and four size-D batteries in his camper. Suppose that the probability than any particular battery works is $p$ and that batteries work or fail independently of one another. Our friend wants to take just one flashlight. For what values of $p$ should he take the 6-V flashlight?**

**Problem: A $k$-out-of-$n$ system is one that functions if and only if at least $k$ of the $n$ individual components in the system function. If individual components function independently of one another, each with probability 0.9, what is the probability that a $3$-out-of-$5$ system functions?**

**How many components do you expect to work in a $3$-out-of-$5$ system?**

**Problem: Suppose that only 20% of all drivers come to a complete stop at an intersection having flashing red lights in all directions when no other cars are visible. What is the probability that, of 20 randomly chosen drivers coming to an intersection under these conditions,**

**(a) at most six will come to a complete stop?**

**(b) exactly six will come to a complete stop?**

**(c) at least six will come to a complete stop?**

**(d) How many of the next 20 drivers do you expect to come to a complete stop?**

**Problem: A baseball player with a 300 average has 600 at-bats (attempts) in a season.**

**(a) Propose a pmf for the number of hits $X$.**

**(b) Is the probability distribution reasonable?**

**(c) What is the expected number of hits?**

**Definition: A rv $X$ has a Poisson($\lambda$) distribution, $\lambda > 0$, if it has pmf**

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots$$

**The Poisson distribution is especially good at modelling rare events (more later).**

**For** $X \sim \text{Poisson}(\lambda)$

- $\text{E}(X) = \lambda$
- $\text{V}(X) = \lambda$

## Proof (first result):

$$\begin{aligned}
\text{E}(X) &= \Sigma_{x=0}^{\infty}\ x\ \lambda^x e^{-\lambda}/x! \\
&= e^{-\lambda}\ \Sigma_{x=1}^{\infty}\ x\ \lambda^x/x! \\
&= e^{-\lambda}\lambda\ \Sigma_{x=1}^{\infty}\ \lambda^{x-1}/(x-1)! \\
&= e^{-\lambda}\lambda\ \Sigma_{y=0}^{\infty}\ \lambda^y/y! \\
&= e^{-\lambda}\lambda\ e^{\lambda} \\
&= \lambda
\end{aligned}$$

**Time for some fun (not related to Poisson)!**

**15 + 6**

**3 + 56**

**89 + 2**

**53 - 12**

**75 + 26**

**-7 - 9**

**123 + 5**

**Quick: Think about a __________ and a __________.**

# Lecture 15

**Recall that the Binomial pmf may be difficult to calculate. It turns out the Binomial can sometimes be approximated by the Poisson.**

**Proposition: Without being rigorous,** $\text{Bin}(n, \theta) \approx \text{Poisson}(n\theta)$ **if** $n$ **is much larger than** $n\theta$**.**

**Example: A rare type of blood occurs in a population with frequency 0.001. If $n$ people are tested, what is the probability that at least two people have this rare blood type? Calculate the probability using the Binomial distribution and the Poisson approximation to the Binomial.**

**Example: A shipment of 5000 parts arrives where 0.5% of the parts are nonconforming. We randomly select 25 parts from the shipment and we reject the entire shipment if more than three of the selected parts are nonconforming. What is the probability that the shipment is accepted?**

**Recall that the Binomial distribution can be motivated by considering $n$ independent trials where the probability of success on each trial is constant. Similarly, the Poisson distribution can be motivated by three assumptions which comprise the *Poisson process*. The assumptions of the Poisson process are these:**

1. **events are indpt in non-overlapping intervals**

2. **events are *stationary***

3. **during small time intervals, the probability of a single event is proportional to the length of the time interval and the probability of more than one event is negligible**

**Proposition: Let $p(x,t)$ be the probability of $x$ successes in an interval of length $t$. Under the assumptions of the Poisson process**

$$p(x,t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \qquad x = 0, 1, \ldots$$

**Example: A switchboard receives calls at a rate of three per minute during a busy period. Let $X_t$ denote the number of calls in $t$ minutes during a busy period. Assess whether the assumptions of a Poisson process are reasonable. Then, assuming the assumptions are reasonable, calculate the probability of receiving more than three calls in a two-minute interval during a busy period.**

# Lecture 16

**Review problem: A limousine can accommodate up to four passengers. The company accepts up to six reservations and passengers must have a reservation to travel. From records, 20% of passengers with reservations do not show.**

**(a) If six reservations are made, what is the probability that at least one passenger cannot be accommodated?**

**(b) If six reservations are made what is the expected number of available places when the limousine departs?**

**(c) Suppose that the pmf of the number of reservations $R$ is**

| $r$ | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| $p(r)$ | 0.1 | 0.2 | 0.3 | 0.4 |

**Find the pmf of the number of passengers $X$ who show up.**

**Definition: A rv is *continuous* if it takes on real values in an interval.**

**Example: Let $X$ be the temperature in degrees Celsius at SFU.**

**Definition: Let $X$ be a continuous rv. Then the *probability density function* (pdf) $f(x) \geq 0$ of $X$ is such that**

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx \qquad \forall a < b$$

**Proposition: The function $f(x)$ is a pdf if**

**1.** $f(x) \geq 0$ **and**

**2.** $\int_{-\infty}^{\infty} f(x)\ dx = 1$

**Problem: Verify that $f(x)$ is a pdf where**

$$f(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 < x \leq 1 \\ 1/2 & 1 < x \leq 2 \\ 0 & 2 < x \end{cases}$$

**Calculate** $\text{Prob}(1 \leq X \leq 1.5)$**.**

**Definition: A rv $X$ has a Uniform$(a, b)$ distribution if it has pdf**

$$f(x) = \frac{1}{b-a} \qquad a < x < b$$

**Special case: Uniform(0,1)**

**Definition: The *cumulative distribution function* (cdf) of a continuous rv $X$ is given by**

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(y)\ dy \qquad a < x < b$$

**Definition: The 100$p$-th *percentile* of the continuous distribution with cdf $F(x)$ is the value $\eta(p)$ such that**

$$p = F(\eta(p))$$

**Definition: The *median* $\tilde{\mu}$ of the continuous distribution with cdf $F(x)$ is the 50-th percentile (i.e. $0.5 = F(\tilde{\mu})$).**

**Example: Find the median of the** Uniform$(a, b)$ **distribution.**

**Definition: The expected value of a continuous rv $X$ with pdf $f(x)$ is**

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)\ dx \qquad a < x < b$$

**Proposition: If $X$ is a continuous rv with pdf $f(x)$**

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\ dx$$

**Definition: The variance of a continuous rv $X$ with pdf $f(x)$ is**

$$\mathrm{V}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x)\ dx$$

**Proposition: If $X$ is a continuous rv, then as in the discrete case,**

- $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$
- $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$
- $\mathrm{V}(aX + b) = a^2\mathrm{V}(X)$

# Lecture 17

**Problem: Consider the pdf of the rv $Y$**

$$f(y) = \begin{cases} y/25 & 0 \leq y < 5 \\ 2/5 - y/25 & 5 \leq y < 10 \end{cases}$$

**(a) obtain the cdf of $Y$**

**(b) calculate the 100$p$-th percentile of $Y$**

**(c) calculate $\mathrm{E}(Y)$**

**Problem: Let $X$ be the time in hours that a reserved book is checked out by a randomly selected student. Suppose that $X$ has the density function**

$$f(x) = \begin{cases} x/2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

**(a) calculate** $\mathrm{P}(X \leq 1)$

**(b) calculate** $\mathrm{P}(0.5 \leq X \leq 1.5)$

**(c) calculate** $\mathrm{P}(0.5 < X)$

**Problem: A professor never finishes lectures before the end of the hour and always finishes within two minutes after the hour. Let $X$ be the time that elapses between the end of the hour and the end of the lecture and suppose the pdf of $X$ is**

$$f(x) = \begin{cases} kx^2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

**(a) evaluate $k$**

**(b) what is the probability that the lecture ends within one minute of the end of the hour?**

**(c) what is the probability that the lecture continues beyond the hour for between 60 and 90 seconds?**

**(d) what is the probability that the lecture continues for at least 90 seconds beyond the end of the hour?**

**Problem: The cdf of checkout duration in minutes $X$ is**

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/4 & 0 \leq x \leq 2 \end{cases}$$

**(a) calculate** $\mathrm{P}(0.5 \leq X \leq 1)$

**(b) calculate the median of** $X$

**(c) calculate the pdf of** $X$

**(d) calculate** $\mathrm{E}(X)$

**Without doubt, the most important distribution in all of Statistics is the normal (Gaussian) distribution.**

**Definition: A rv $X$ has a Normal$(\mu, \sigma^2)$ distribution if it has pdf**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

**where $x \in \mathcal{R}$, $\mu \in \mathcal{R}$ and $\sigma > 0$.**

**Some talking points:**

- **the normal is a family of distributions**
- **the density is symmetric about $\mu$**
- **the density never touches zero**
- **the density is not tractable**
- **the parameters are interpretable: $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \sigma^2$**
- **data are often approximately normal**
- **the standard normal distribution is $\mathrm{Normal}(0, 1)$ and is typically represented by the rv $Z$**

**To gain an understanding of the parameters $\mu$ and $\sigma$, sketch plots of the following distributions:**

- $\mathrm{Normal}(5, 1)$
- $\mathrm{Normal}(7, 1)$
- $\mathrm{Normal}(5, 10)$
- $\mathrm{Normal}(5, 1/10)$

**You must become familiar with the standard normal table (Table B.2 in the text). Calculate the following:**

**(a)** $\mathrm{P}(Z \leq 3.02)$

**(b)** $\mathrm{P}(Z > 3.03)$

**(c)** $\mathrm{P}(Z < 3.025)$ **via interpolation**

**(d)** $\mathrm{P}(2.3 \leq Z \leq 2.6)$

**(e)** $\mathrm{P}(Z > -1)$

**(f)** $z$ **such that 30.5% of** $Z$**-values exceed** $z$

# Lecture 18

**Proposition: If $X \sim \mathrm{Normal}(\mu, \sigma^2)$, then**

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{Normal}(0, 1)$$

**A consequence is that any normal probability can be converted to a probability involving the standard normal. This means that we only need a single normal table instead of tables for all possible values of $\mu$ and $\sigma$.**

**Problem: The number of hours that people watch television is normally distributed with mean 6.0 hours and standard deviation 2.5 hours (first ask yourself if this is reasonable). What is the probability that a randomly selected person watches more than 8 hours of television per day?**

**Problem: The substrate concentration (mg/cm$^3$) of influent to a reactor is normally distributed with $\mu = 0.30$ and $\sigma = 0.06$.**

**(a) What is the probability that the concentration exceeds 0.25?**

**(b) What is the probability that the concentration is at most 0.10?**

**(c) How would you characterize the largest 5% of all concentration values?**

**An amusing but real problem: My wife was expecting on June 1. My friends wanted me to go on a golf trip May 14, 15 and 16. What to do?**

**Proposition: Let $\eta(p)$ denote the $100p$-th percentile of the standard normal distribution. Then the $100p$-th percentile of the $\mathrm{Normal}(\mu, \sigma^2)$ distribution is $\mu + \sigma\eta(p)$.**

**Example: Find the 25.78-th percentile of the $\mathrm{Normal}(5, 100)$.**

**Proposition: Consider** $X \sim \mathrm{Bin}(n, p)$ **where** $np \geq 5$ **and** $n(1-p) \geq 5$. **Then we have the following approximation**

$$X \sim \mathrm{Normal}(np, np(1-p))$$

**Example: Obtain** $\mathrm{P}(X \geq 8)$ **where** $X \sim \mathrm{Bin}(10, 1/2)$

**(a) exactly**

**(b) using the normal approximation**

**(c) using the normal approximation with a *continuity correction*.**

**Reminders:**

**(1) Probabilities associated with the** $\mathrm{Bin}(n,p)$ **are sometimes difficult to evaluate. The following approximations are available:**

- **(a)** $\mathrm{Poisson}(np)$ **if** $n$ **is large and** $p$ **is small**
- **(b)** $\mathrm{Normal}(np, np(1-p))$ **if** $np \geq 5$ **and** $n(1-p) \geq 5$

**(2) Use a continuity correction whenever you need to approximate a discrete distribution with a continuous distribution.**

# Lecture 19

**Problem: Verbal SAT scores are normally distributed with a mean score of 430 and variance 100. What is the middle range of scores encompassing 50% of the population?**

**Problem: The automatic opening device of a military parachute is designed to open when the parachute is 200 metres above ground. Suppose that the opening altitude has a normal distribution with mean 200m and standard deviation 30m. Equipment damage occurs if the parachute opens at an altitude of less than 100m above ground. What is the probability that there is equipment damage to at least one of five independently dropped parachutes?**

**Problem: The temperature reading from a thermocouple in a constant-temperature medium is normally distributed with mean $\mu$ (the actual temperature of the medium) and standard deviation $\sigma$. What is the value of $\sigma$ such that 95% of all readings are within 0.1 degree of $\mu$?**

**Problem: A patient is hypokalemic if their level of potassium is 3.5 or less. An individual's level is not constant, but varies daily. Suppose that the variation is normal. Judy has a mean level of 3.8 with variance 0.04. If she is measured daily, what proportion of days would she be declared hypokalemic?**

**Problem:** A college has a target enrollment of 1200 students. Since not all admitted students actually enroll, the college admits 1500 students. Past experience shows that 70% of students who are offered admission enroll.

(a) Give a statistical model for the number of students who enroll.

(b) Obtain the corresponding mean and std dev.

(c) Obtain the prob that at least 1200 enroll.

**Problem: The volume placed in a bottle by a bottling machine follows a $\text{Normal}(\mu, \sigma^2)$ distribution. Over a long period of time, it is observed that 5% of the bottles contain less than 31.5oz and 15% contain more than 32.3oz.**

**(a) Find $\mu$ and $\sigma$.**

**(b) Calculate the probability that out of 10 bottles purchased, exactly three bottles contain more than 32.2 oz.**

# Lecture 20

**Problem: The weight distribution of parcels is normal with mean value 12lb and std dev 3.5lb. The parcel service wants to establish a weight $c$ beyond which there is a surcharge. What is the value of $c$ such that 99% of parcels are at least 1lb under the surcharge weight?**

**Problem: The breakdown voltage of a randomly chosen diode is normally distributed with mean 40V and standard deviation 1.5V.**

**(a) What is the probability that the voltage of a single diode is between 39V and 42V?**

**(b) What value is such that only 15% of diodes have voltages exceeding that value?**

**(c) If four diodes are randomly selected, what is the probability that at least one has voltage exceeding 42V?**

**Definition: A rv $X$ has a $\mathrm{Gamma}(\alpha, \beta)$ distribution, $\alpha > 0$, $\beta > 0$, if it has pdf**

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \qquad\qquad x > 0$$

**where** $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx$

**Discussion points:**

- **pdf generally intractable**
- **contrast the range (**$x > 0$**) with the normal**
- **asymmetric**
- $\Gamma(\alpha)$ **is a constant**
- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$**,** $\Gamma(1) = 1$**,** $\Gamma(1/2) = \sqrt{\pi}$

**Proposition: If $X \sim \text{Gamma}(\alpha, \beta)$, then**

- $\text{E}(X) = \alpha\beta$
- $\text{V}(X) = \alpha\beta^2$

**The** $\text{Exponential}(\lambda)$ **distribution is a special case of the** $\text{Gamma}(\alpha, \beta)$ **where** $\alpha = 1$ **and** $\beta = 1/\lambda$**.**

**Definition: A rv** $X$ **has an** $\text{Exponential}(\lambda)$ **distribution,** $\lambda > 0$**, if it has pdf**

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0$$

**Discussion points:**

- $\text{E}(X) = \alpha\beta = 1(1/\lambda) = 1/\lambda$
- $\text{V}(X) = \alpha\beta^2 = 1(1/\lambda)^2 = 1/\lambda^2$
- **the density is decreasing for** $x > 0$
- **the density is tractable; in particular the cdf** $F(x) = 1 - e^{-\lambda x}$ **for** $x > 0$

**The Exponential distribution possesses a curious property known as the *memoryless* property. To appreciate the property, consider a rv $X$ which is the lifespan of a lightbulb in hours where we assume that $X \sim \mathrm{Exponential}(\lambda)$. Then the probability that a used lightbulb (that has already lasted $a$ hours) will last an additional $b$ hours is given by**

$$\mathrm{P}(X > a + b \ \mid X > a) =$$

# Lecture 21

**It turns out that there is a connection between the Poisson and Exponential distributions. Recall the Poisson process where $N_T$ is the number of events that occur in the interval $[0, T]$ where $N_T \sim \text{Poisson}(\lambda T)$. Let**

$$Y \equiv \textbf{waiting time until the first event}$$

**Then the cdf of $Y$ is given by**

$$\begin{aligned}
\mathrm{P}(Y \leq y) &= 1 - \mathrm{P}(Y > y) \\
&= 1 - \mathrm{P}(\textbf{zero events in [0,y]}) \\
&= 1 - \mathrm{P}(N_y = 0) \qquad \text{where } N_y \sim \text{Poisson}(\lambda y) \\
&= 1 - (\lambda y)^0 e^{-\lambda y}/0! \\
&= 1 - e^{-\lambda y}
\end{aligned}$$

**which implies $Y \sim \text{Exponential}(\lambda)$**

**Problem: Let $X$ be the distance in metres that a rat moves from its birth site to its first territorial vacancy. Suppose that $X$ has an exponential distribution with $\lambda = 0.01386$.**

**(a) What is the probability that the distance $X$ is at most 100 metres?**

**(b) What is the probability that the distance $X$ exceeds the mean distance by more than two standard deviations?**

**(c) What is the median distance?**

**Until now, we have studied probabilities corresponding to a single rv $X$. We now consider joint probability distributions associated with a vector rv $(X_1, \ldots, X_k)$.**

**Example: a trivariate discrete distribution described by the pmf** $p(x, y, z)$

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.10 | 0.20 | 0.00 |
| **Y=2** | 0.00 | 0.05 | 0.05 |

$Z = 5$

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.00 | 0.30 | 0.10 |
| **Y=2** | 0.05 | 0.05 | 0.10 |

$Z = 6$

**The *marginal* pmf** $\mathrm{p}(x) = \Sigma_{y,z}\mathrm{p}(x, y, z)$

**In the continuous setting, we describe distributions via a joint pdf $f(x_1, \ldots, x_k)$ which satisfies**

**1.** $f(x_1, \ldots, x_k) \geq 0 \qquad \forall\ x_1, \ldots, x_k$

**2.** $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_k)\ dx_1 \cdots dx_k \ = \ 1$

**To obtain probabilities in the continuous setting,**

$\mathrm{P}((X_1, \ldots, X_k) \in A) = \int \cdots \int_A f(x_1, \ldots, x_k)\ dx_1 \cdots dx_k$

**Example: A bivariate distribution on $(X, Y)$ is given by $f(x, y) = 2(2x + 3y)/5$ where $0 < x, y < 1$**

**(a) Calculate $\mathrm{P}(X > 1/2, Y < 1/2)$.**

**(b) Obtain the marginal pdf of $X$ and verify that it is a pdf.**

**Recall that we previously discussed the independence of events. The concept of independence can be extended to rv's.**

**Definition: Random variables are independent if their joint pmfs (pdfs) factor into their marginal pmfs (pdfs).**

**Example: Consider the bivariate pdf**

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left(x^2/\sigma_1^2+y^2/\sigma_2^2\right)\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{1}{2}\left(\frac{x-0}{\sigma_1}\right)^2\right\}\ \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{1}{2}\left(\frac{y-0}{\sigma_2}\right)^2\right\} \end{aligned}$$

**Example: Consider the bivariate pmf given by**

| | **X=1** | **X=2** |
|---|---|---|
| **Y=1** | 0.4 | 0.2 |
| **Y=2** | 0.1 | 0.3 |

**(a) Obtain the marginal pmf for $X$.**

**(b) Obtain the marginal pmf for $Y$.**

**(c) Are $X$ and $Y$ independent?**

**Problem: Two components of a computer have the joint pdf for their lifetimes $X$ and $Y$ in years**

$$f(x, y) = xe^{-x(1+y)} \qquad x, y \geq 0$$

**(a) What is the probability that the lifetime $X$ of the first component exceeds 3 years?**

**(b) What are the marginal pdfs of $X$ and $Y$?**

**(c) What is the probability that the lifetime of at least one component exceeds 3 years?**

# Lecture 22

**We now turn our attention to the expectation of functions of random variables.**

**Proposition: In the continuous case, using standard notation,**

$$\mathrm{E}[g(X_1,\ldots,X_k)] = \int\cdots\int g(x_1,\ldots x_k)f(x_1,\ldots x_k)\ dx_1\cdots dx_k$$

**In the discrete case, we replace the pdf $f$ with the corresponding pmf $p$ and we replace the multiple integral with a multiple sum.**

**Example: An instructor gives a quiz with two parts. For a randomly selected student, let $X$ and $Y$ be the scores obtained on the two parts respectively. The table gives the joint pmf $p(x, y)$ of $X$ and $Y$:**

| **p(x,y)** | **y=0** | **y=5** | **y=10** | **y=15** |
|---|---|---|---|---|
| **x=0** | 0.02 | 0.06 | 0.02 | 0.10 |
| **x=5** | 0.04 | 0.15 | 0.20 | 0.10 |
| **x=10** | 0.01 | 0.15 | 0.14 | 0.01 |

**(a) What is the expected total score $\mathrm{E}(X + Y)$?**

**(b) What is the expected maximum score from the two parts?**

**(c) Are $X$ and $Y$ independent?**

**(d) Obtain $\mathrm{P}(Y = 10 \mid X \geq 5)$.**

**Example: We return to the discrete distribution described by the pmf** $p(x, y, z)$

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.10 | 0.20 | 0.00 |
| **Y=2** | 0.00 | 0.05 | 0.05 |

$Z = 5$

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.00 | 0.30 | 0.10 |
| **Y=2** | 0.05 | 0.05 | 0.10 |

$Z = 6$

**Obtain** $\mathrm{E}(g)$ **where** $g(x, y, z) = xz$**.**

**Problem: Annie and Alvie agree to meet for lunch between noon and 1pm. Denote Annie's arrival time by $X$ and Alvie's by $Y$, and suppose $X$ and $Y$ are independent with pdfs $f_X(x) = 3x^2$ where $0 < x < 1$ and $f_Y(y) = 2y$ where $0 < y < 1$.**

**What is the expected time that the one who arrives first waits for the other person to arrive?**

**Recall that the conditional probability of event $A$ given event $B$ is given by**

$$\mathrm{P}(A \mid B) = \mathrm{P}(AB)/\mathrm{P}(B).$$

**Conditional probability is now extended to continuous rv's.**

**Definition: In the continuous case, using standard notation, the conditional density of $X_1$ given $X_2 = x_2, \ldots, X_k = x_k$ is given by**

$$f_{X_1|X_2,\ldots,X_k}(x_1) = \frac{f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}{f_{X_2,\ldots,X_K}(x_2, \ldots, x_k)}$$

**The definition can be extended in various ways including the discrete case.**

**Example: Recall the bivariate distribution on** $(X, Y)$ **given by the pdf** $f_{X,Y}(x, y) = 2(2x + 3y)/5$ **where** $0 < x, y < 1$**. Earlier we established the marginal density for** $X$ **given by** $f_X(x) = 4x/5+3/5$ **where** $0 < x < 1$**. Suppose we observe** $X = 0.2$**. What is the conditional pdf of** $Y$**?**

**Problem: The number of customers waiting for the gift-wrap service at department store is a rv $X$ taking possible values 0, 1, 2, 3 and 4 with corresponding probabilities 0.10, 0.20, 0.30, 0.25 and 0.15. A random customer has 1, 2 or 3 packages for wrapping with probabilities 0.6, 0.3 and 0.1 respectively. Let $Y$ be the total number of packages to be wrapped by customers waiting in line.**

**(a) Determine** $\mathrm{P}(X = 3, Y = 3)$**.**

**(b) Determine** $\mathrm{P}(X = 4, Y = 11)$**.**

# Lecture 23

**Definition: The *covariance* between the rvs $X$ and $Y$ is given by**

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}(\ (X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\ ) \\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)\end{aligned}$$

**Interpretation:**

- **positive covariance**
  - **large $x$'s occur with large $y$'s**
  - **small $x$'s occur with small $y$'s**
- **negative covariance**
  - **large $x$'s occur with small $y$'s**
  - **small $x$'s occur with large $y$'s**

**Correlation is the scaled and preferred version of covariance.**

**Definition: The *correlation* between the rvs $X$ and $Y$ is given by**

$$\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}}$$

**Discussion points:**

- $-1 \leq \mathrm{Corr}(X, Y) \leq 1$
- **correlation is location/scale invariant**
- $\rho$ **is the population analogue of** $r$
- $\rho$ **typically relevant to continuous rvs**
- **if** $a > 0$**, then** $\mathrm{Corr}(X, aX + b) = 1$
- **if** $a < 0$**, then** $\mathrm{Corr}(X, aX + b) = -1$

**Example: Obtain the correlation between $X$ and $Y$ where the joint pmf of $X$ and $Y$ is given in the following table.**

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.1 | 0.2 | 0.3 |
| **Y=2** | 0.0 | 0.2 | 0.2 |

**Proposition: If $X$ and $Y$ are independent, then**

$$\mathrm{Cov}(X, Y) = 0$$

**In addition,** $\mathrm{Corr}(X, Y) = 0$ **provided** $\mathrm{V}(X)$ **and** $\mathrm{V}(Y)$ **are nonzero. The converse is not true.**

**Also, recall that correlation does not imply causation.**

**Proposition:** $\mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{Cov}(X,Y)$

**Proposition: More generally,**

$$\mathrm{V}(aX + bY + c) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Y) + 2ab\mathrm{Cov}(X,Y)$$

**Proposition: Even more generally,**

$$\mathrm{V}\left(\sum_{i=1}^{n} a_i X_i + c\right) = \textstyle\sum_{i=1}^{n} a_i^2 \mathrm{V}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i + c\right) = \textstyle c + \sum_{i=1}^{n} a_i \mathrm{E}(X_i)$$

**Lets put some of this stuff together to provide a useful result.**

**Corollary: Suppose that the rv's $X_1, \ldots, X_n$ are a *sample*. In other words, the $X$'s are independent and arise from a common distribution with mean $\mu$ and variance $\sigma^2$. Then the sample mean has the following properties:**

- $\mathrm{E}(\bar{X}) = \mu$
- $\mathrm{V}(\bar{X}) = \sigma^2/n$

**Suprisingly, we have reached this point in our Statistics course and we have not yet defined the word *statistic*.**

**Definition: A *statistic* is a function of the data.**

**Some examples:**

- $\bar{X} = \sum_{i=1}^{n} X_i/n$ **is a statistic**
- $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ **is a statistic**

**Since data are variable, statistics are also variable. Sometimes we are interested in the distributions of statistics.**

**Example: Obtain the distribution of the statistic $Q = X+Y$ where the joint pmf of $X$ and $Y$ is given in the following table.**

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.1 | 0.1 | 0.2 |
| **Y=2** | 0.2 | 0.3 | 0.1 |

**The previous example was simple. To generalize, we need to go a little crazy with notation.**

**Suppose that $X_1, \ldots, X_n$ are discrete with joint pmf $p(x_1, \ldots, x_n)$. Then the pmf for the general statistic $Q(X_1, \ldots, X_n)$ is**

$$p_Q(q) = \sum_A p(x_1, \ldots, x_n)$$

**where the sum is a multiple sum and $A$ is the set of $x_1, \ldots, x_n$ such that $Q(x_1, \ldots, x_n) = q$.**

**Suppose that $X_1, \ldots, X_n$ are continuous with joint pdf $f(x_1, \ldots, x_n)$. Then the cdf for the general statistic $Q(X_1, \ldots, X_n)$ is**

$$F_Q(q) = \mathrm{P}(Q \leq q) = \int_A f(x_1, \ldots, x_n)\ dx_1 \ldots dx_n$$

**where the integral is a multiple integral and $A$ is the set of $x_1, \ldots, x_n$ such that $Q(x_1, \ldots, x_n) \leq q$.**

**I have mentioned previously that statistical practice relies heavily on computation. Here is a simulation procedure that can be used to approximate distributions of statistics when the sums and integrals from the previous page are too difficult to obtain analytically.**

- **Repeat the following two steps $M$ times where $M$ is large and let $i$ denote the $i$-th iteration**
  - **generate $x_1, \ldots, x_n$ according to $p(x_1, \ldots, x_n)$ or $f(x_1, \ldots, x_n)$ (depending whether the data are discrete or continuous)**
  - **calculate $Q_i = Q(x_1, \ldots, x_n)$ for the data**
- **approximate the distbn of Q with a histogram based on generated outcomes $Q_1, \ldots, Q_M$**

# Lecture 24

**Proposition: Linear combinations of normal rv's are normal.**

**Corollary: Suppose that $X_1, \ldots, X_n$ is a sample from the $\mathrm{Normal}(\mu, \sigma^2)$ distribution. Then**

$$\bar{X} \sim \mathrm{Normal}(\mu, \sigma^2/n)$$

**Example: Determine the distribution of the rv** $Y = 2X_1 - X_2 + 3X_3 + 3$ **where** $X_1$**,** $X_2$ **and** $X_3$ **are independent,** $X_1 \sim \text{Normal}(4, 3)$**,** $X_2 \sim \text{Normal}(5, 7)$ **and** $X_3 \sim \text{Normal}(6, 4)$**.**

**Example: Determine the distribution of the rv** $Y = X_1 - X_2$ **where** $\text{Cov}(X_1, X_2) = 6$**,** $X_1 \sim \text{Normal}(5, 10)$ **and** $X_2 \sim \text{Normal}(3, 8)$**.**

**You are not responsible for complete understanding of the following example. However, it gives some insight as to why linear combinations of normals are normal.**

**Example: When $X$ and $Y$ are independent standard normal, then $Z = X + Y \sim \mathrm{Normal}(0,2)$.**

$$\begin{aligned}
\mathrm{P}(Z \le z) &= \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{z-y} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\ dx\ dy \\
&= \int_{y=-\infty}^{\infty}\int_{u=-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-(u-y)^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\ du\ dy \\
&= \int_{u=-\infty}^{z}\int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}e^{-u^2/2+uy-y^2}\ dy\ du \\
&= \int_{u=-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-(y^2-uy)}\ dy\ du \\
&= \int_{u=-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-(y-u/2)^2+u^2/4}\ dy\ du \\
&= \int_{u=-\infty}^{z} \frac{\sqrt{1/2}}{\sqrt{2\pi}}e^{-u^2/4}\int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1/2)}}e^{-\frac{1}{2}\left(\frac{y-u/2}{\sqrt{1/2}}\right)^2}\ dy\ du \\
&= \int_{u=-\infty}^{z} \frac{1}{\sqrt{2\pi(2)}}e^{-\frac{1}{2}\left(\frac{u}{\sqrt{2}}\right)^2}\ du
\end{aligned}$$

**Problem: Suppose that the waiting time for a bus in the morning is uniformly distributed on [0,8] whereas the waiting time for a bus in the evening is uniformly distributed on [0,10]. Assume that the waiting times are independent.**

**(a) If you take a bus each morning and evening for a week, what is the total expected waiting time?**

**(b) What is the variance of total waiting time?**

**(c) What are the expected value and variance of how much longer you wait in the evening than in the morning on a given day?**

**Proposition - The Central Limit Theorem (CLT): Let $X_1, \ldots, X_n$ be iid (independent and identically distributed) rvs arising from a distribution with mean $\mu$ and variance $\sigma^2$. Then as $n \to \infty$,**

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathrm{Normal}(0, 1)$$

**Discussion points:**

- **the most important (and arguably) most beautiful result in Statistics**
- **weaker versions of the CLT are available**
- **the CLT is important for inference**
- **assuming little, the CLT tells us a lot**
- **try to understand the limits used in the CLT**
- **we use the limiting distribution when the sample size is large ($n \geq 30$)**

**We motivate the CLT by considering a sample $X_1, \ldots X_n$ with underlying pmf $p(x)$.**

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $p(x)$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ |

**(a) Obtain the distribution of $\bar{X}$ when $n = 2$.**

**(b) Obtain the distribution of $\bar{X}$ when $n = 3$.**

# Lecture 25

**Example: Suppose that you order 500 apples and you know from previous orders that the mean weight of an apple is 0.2 kg with std dev 0.1 kg. What is the probability that the total weight of the 500 apples is less than 98 kg?**

**Problem: A restaurant serves three dinners costing \$12, \$15 and \$20. For a randomly selected couple, let $X$ be the cost of the man's dinner and let $Y$ be the cost of the woman's dinner. The joint pmf of $X$ and $Y$ is given as shown.**

|  |  | $y$ |  |  |
|---|---|---|---|---|
| $p(x,y)$ |  | 12 | 15 | 20 |
|  | 12 | .05 | .05 | .10 |
| $x$ | 15 | .05 | .10 | .35 |
|  | 20 | .00 | .20 | .10 |

**(a) Suppose that when a couple opens a fortune cookie, they find the message "You receive a refund equal to the difference between the cost of your most expensive and least expensive meal". How much does the restaurant expect to refund?**

**Problem: I have three errands where $X_i$ is the time required for the $i$-th errand, $i = 1, 2, 3$ and $X_4$ is the total walking time between errands. Suppose that the $X$'s are independent normal rvs with means $\mu_1 = 15$, $\mu_2 = 5$, $\mu_3 = 8$, $\mu_4 = 12$, and standard deviations $\sigma_1 = 4$, $\sigma_2 = 1$, $\sigma_3 = 2$, $\sigma_4 = 3$. I plan to leave my office at 10 am and post a note on the door reading "I will return by $t$ am."**

**(a) What time $t$ ensures that the probability of arriving later than $t$ is 0.01?**

**Problem: The mean tensile strength of type-A steel is 105 ksi with standard deviation 8 ksi. For type-B steel, the mean tensile strength is 100 ksi and standard deviation 6 ksi. Let $\bar{X}$ be the sample average of 40 type-A specimens and let $\bar{Y}$ be the sample average of 35 type-B specimens.**

**(a) What are the approx distbns of $\bar{X}$ and $\bar{Y}$?**

**(b) What is the approx distbn of $\bar{X} - \bar{Y}$?**

**(c) Calculate approximately $\mathrm{P}(-1 \le \bar{X} - \bar{Y} \le 1)$.**

**Problem: Let $X_1, \ldots, X_n$ be rvs corresponding to $n$ independent bids for an item on sale. Suppose each $X_i$ is uniformly distributed on $[100, 200]$.**

**(a) If the seller sells to the highest bidder, what is the expected sale price?**

**Problem: The mean weight of luggage for an economy passenger is 40 lb with std dev 10 lb. The mean weight of luggage for a business class passenger is 30 lb with std dev 6 lb. Suppose that there are 12 business class and 50 economy passengers on a given flight.**

**(a) What is the expected total luggage weight and standard deviation?**

**(b) What is the prob that the total luggage weight is at most 2500 lb if luggage weights are independent and normally distributed?**

**Problem: If the amount of soft drink I consume is independent of consumption on other days and is normally distributed with $\mu = 13$ oz and $\sigma = 2$ oz, and I currently have two six-packs of 16-oz bottles, what is the probability that I will have some soft drink remaining after two weeks?**

**Problem: In an area with sandy soil, 50 small trees of a certain type are planted, and another 50 trees are planted in an area with clay soil. Let $X$ be the number of surviving trees after one year planted in the sandy soil and let $Y$ be the number of surviving trees after one year planted in the clay soil. Suppose the one-year survival probability of a tree planted in sandy soil is 0.7 and the one-year survival probability of a tree planted in clay soil is 0.6.**

**(a) Approximate** $\mathrm{P}(-5 \leq X - Y \leq 5)$**.**

**Problem: Suppose calorie intake at breakfast is a rv with mean 500 and std dev 50, calorie intake at lunch is a rv with mean 900 and std dev 100, and calorie intake at dinner is a rv with mean 2000 and std dev 180. Assuming that intakes at the three meals are independent, what is the probability that the average daily intake over the next year is at most 3500?**

# Lecture 26

**Our attention now turns to *statistical inference* where we try to understand poplns based on sample data. We first study *confidence intervals.***

**The Problem: Given a statistical model (eg. $X \sim \text{Normal}(\mu, \sigma^2)$, $Y \sim \text{Bin}(n, p)$, $W \sim \text{Poisson}(\theta)$), the estimation problem is to learn about unknown parameters (eg. $\mu$, $\sigma$, $p$, $\theta$) given observed data (eg. $X$'s, $Y$'s, $W$'s).**

**Idea 1: We might estimate the population mean $\mu$ with the *point estimate* $\bar{X}$. Point estimation is barely mentioned in the text. Although seemingly sensible, the problem is that we do not know about the closeness of the estimate $\bar{X}$ to the unknown parameter $\mu$.**

**Idea 2: Interval estimation involves constructing an interval (eg. (7.3,12.6) ) in which we are confident that $\mu$ resides.**
**We begin with confidence interval construction**

**in the simplest context. Consider** $X_1, \ldots, X_n$ **iid** $\text{Normal}(\mu, \sigma^2)$ **where** $\mu$ **is unknown,** $\sigma$ **is known and the observed value of** $\bar{X}$ **is** $\bar{X}_{\text{obs}}$**.**

**Note that this is an unrealistic scenario. When is it ever the case that the mean parameter is unknown but the variance parameter is known? Ignoring the criticism,** $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$**. A 95% confidence interval for** $\mu$ **is obtained via:**

$$\begin{aligned}
&\text{P}\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \\
\Leftrightarrow\ &\text{P}\left(-1.96\frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < 1.96\frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0.95 \\
\Leftrightarrow\ &\text{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \\
\Rightarrow\ &\bar{X}_{\text{obs}} \pm 1.96\frac{\sigma}{\sqrt{n}} \quad \textbf{is a 95\% CI for } \mu
\end{aligned}$$

**More generally,**

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \textbf{is a } (1-\alpha)100\% \textbf{ CI for } \mu.$$

**Interpretation of CI's: The explanation is subtle and you need to pay close attention.**

**Consider many hypothetical replications of an experiment.**

**A common but incorrect interpretation for CI'S:**

**If** $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ **is a** $(1-\alpha)100\%$ **CI for** $\mu$**, it is incorrect to write** $\mathrm{P}\left(\mu \in \bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$**.**

**Discussion points wrt the CI $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$:**

- **as $n$ increases, the width of the CI decreases**
- **as our confidence increases (ie. $1 - \alpha$ bigger), the width of the CI increases**
- **tradeoff: we want narrow CI's with large confidence**
- **a CI of a given confidence $1 - \alpha$ is not unique**

**The simple but unrealistic CI setting previously presented is extended to more realistic scenarios.**

**We begin by assuming that our sample $X_1, \ldots, X_n$ is large (ie. $n \geq 30$) as is often the case in practice.**

**Case 1: Since $n$ is large, we can invoke the CLT where approximately $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$. What is great about this is that we no longer need to assume that the $X$'s are normal. In this case,**

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

**is an approximate $(1-\alpha)100\%$ CI for $\mu$ where $\sigma$ is still assumed known.**

**Case 2: We have the same conditions as Case 1 except that $\sigma$ is unknown. In this realistic case,**

$$\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

**is an approximate $(1-\alpha)100\%$ CI for $\mu$ where $s$ is the sample standard deviation.**

**Example: Consider heat measurements taken in degrees Celsius where $\mu = 5$ and $\sigma = 4$. A change is made in the process such that $\mu$ changes but $\sigma$ remains the same. We observe $\bar{X}_{\text{obs}} = 6.1$ based on $n = 100$ observations.**

**(a) Construct a 90% CI for $\mu$.**

**(b) How big should the sample size be such that the CI is less than 0.6 degrees wide?**

**Problem: Consider the CI $\bar{X}_{\text{obs}} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.**

**(a) How much should the sample size $n$ increase to reduce the width of by half?**

**(b) What is the effect of increasing the sample size by a factor of 25?**

# Lecture 27

**We now construct a confidence interval for the unknown $p$ in the model $X \sim \text{Binomial}(n,p)$. We require $np \geq 5$ and $n(1-p) \geq 5$ (ie. $n$ large and $p$ moderate) so that we can use the approximation $X \sim \text{Normal}(np, np(1-p))$. We denote $\hat{p} = X/n$ and $\hat{p}_{\text{obs}} = X_{\text{obs}}/n$. A $(1-\alpha)$% confidence interval for $p$ is obtained via:**

$$\begin{aligned}
& \mathrm{P}\left(-z_{\frac{\alpha}{2}} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha \\
\Leftrightarrow\ & \mathrm{P}\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha \\
\Leftrightarrow\ & \mathrm{P}\left(-z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} - \hat{p} < -p < z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} - \hat{p}\right) = 1-\alpha \\
\Leftrightarrow\ & \mathrm{P}\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}\right) = 1-\alpha
\end{aligned}$$

**Therefore,**

$$\hat{p}_{\text{obs}} \pm z_{\frac{\alpha}{2}}\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n} \qquad (1)$$

**is an approximate $(1-\alpha)100\%$ CI for $p$. The CI (1) is based on two approximations:**

1. **approximating the Binomial with the Normal**
2. **replacing $p$ with $\hat{p}$**

**Example: From a sample of 1250 BC voters, 420 indicate that they support the NDP. Obtain an approximate 95% CI for the proportion of BC voters who support the NDP.**

**CI's based on the Student distribution: Suppose $X_1, \ldots, X_n$ are iid $\text{Normal}(\mu, \sigma^2)$ where $\sigma$ is unknown (the realistic case). It can be shown that**

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

**where $t_{n-1}$ denotes the $t$ distribution with $n-1$ degrees of freedom. The pdf of $Y \sim t_{n-1}$ is**

$$f(y) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)\Gamma(1/2)}\left(1 + \frac{y^2}{n-1}\right)^{-n/2}$$

**Here, the $(1-\alpha)$100% confidence interval for $\mu$ is**

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

**Discussion points:**

- **the $t_{n-1}$ distribution is symmetric on $\mathcal{R}$**
- **the $t_{n-1}$ has longer tails than the normal**
- **as $n \to \infty$, $t_{n-1} \to Z \sim \text{Normal}(0,1)$**
- **for $n \geq 30$, you can replace $t_{n-1}$ with $Z$**
- **the $t$ distribution is intractable; no need to memorize pdf**
- **Table B.1 in the text gives points $t_{\frac{\alpha}{2},n-1}$**

**The logic of hypothesis testing: We view the testing of hypotheses as consisting of three steps. We discuss the three steps in some detail.**

1. **The experimenter forms a *null hypothesis* $\mathrm{H}_0$ to test against an *alternative hypothesis* $\mathrm{H}_1$.**
2. **The experimenter collects data.**
3. **In the inference step, the question is asked "Are the data compatible wrt $\mathrm{H}_0$?" If yes, $\mathrm{H}_0$ is not rejected. If no, $\mathrm{H}_0$ is rejected.**

**Example: In this informal example, we go over the three steps of hypothesis testing. Imagine a court of law where a defendent is accused of a crime.**

**Example: In this informal example, we go over the three steps of hypothesis testing. Imagine that you are playing cards and that your friend has obtained a royal flush three hands in a row.**

**In the inference step, if we answer “yes” to the key question (Are the data compatible wrt $\mathrm{H}_0$?), we conclude using the curious language, “$\mathrm{H}_0$ is not rejected”. We discuss why this does not mean the same thing as “$\mathrm{H}_0$ is accepted”.**

# Lecture 28

**In this lecture, we examine five examples each of which does something different in the context of hypothesis testing.**

**To address the inference step (step 3 of hypothesis testing), we compute a *p-value* which is defined as the probability of observing a result as extreme or more extreme than what we observed given that $\mathrm{H}_0$ is true (think about this!)**

**The convention is to reject $\mathrm{H}_0$ and conclude $\mathrm{H}_1$ if the p-value is less than 0.05. Sometimes a stronger level of evidence is required (e.g. 0.01).**

**Example: A shop sells coffee where the number of lb of coffee sold in a week is** $\mathrm{Normal}(320, 40^2)$**. After advertising, 350 lb is sold in the following week. Has advertising improved business?**

**Example: A soup company makes soup in 10 oz cans. A sample of 48 cans has mean volume 9.82 oz and $s = 0.8$ oz. Can we conclude that the company is cheating? Test at level 0.01 significance.**

**Example: A coin is flipped 10 times and 8 heads appear. Is the coin fair?**

**Example: A coin is flipped 100 times and 60 heads appear. Is the coin fair?**

**Example: A paint is applied to tin panels and baked for one hour such that the mean index of hardness is 35.2. Suppose 20 panels are painted and baked for three hours, and their sample mean index of hardness is 37.2 with** $s = 1.4$**. Does baking for three hours strengthen panels? Assume normal data.**

# Single Sample Testing - Summary

| Data | **Test Statistic** | **Comments** |
|---|---|---|
| **normal**,<br>$\sigma$ **known** | $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ | unrealistic |
| **normal**,<br>$\sigma$ **unknown** | $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ | **N(0,1) when** $n \geq 30$ |
| **non-specified**,<br>$\sigma$ **known,**<br>$n \geq 30$ | $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ | unrealistic<br>**based on CLT** |
| **non-specified**,<br>$\sigma$ **unknown,**<br>$n \geq 30$ | $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim \mathrm{N}(0,1)$ | **based on CLT** |
| Binomial | Binomial | |
| **Binomial**,<br>$np \geq 5$**,**<br>$n(1-p) \geq 5$ | $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim \mathrm{N}(0,1)$ | |

# Lecture 29

**We now study the two sample problem where the data $X_1, \ldots, X_m$ iid $\mathrm{Normal}(\mu_1, \sigma_1^2)$ is independent of $Y_1, \ldots, Y_n$ iid $\mathrm{Normal}(\mu_2, \sigma_2^2)$. Initially, we make the unrealistic assumption that both $\sigma_1$ and $\sigma_2$ are known.**

**Under the above conditions, interest lies in the unknown parameter $\mu_1 - \mu_2$. The test statistic used in the construction of confidence intervals and hypothesis testing is**

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{Normal}(0, 1)$$

**Example: Suppose that fifty test scores from class A are independent of 70 test scores from class B. Assume further that the test scores are normal, $\sigma_1^2 = \sigma_2^2 = 84$, $\bar{X} = 73$ and $\bar{Y} = 59$. Is there a difference between the two classes?**

**Example continued: Construct a 95% confidence interval for $\mu_1 - \mu_2$.**

**Example continued: Suppose the question had instead been, "Is class A better than class B?"**

**Example continued: Suppose the question had instead been, “Is class A more than five marks better than class B?”**

**The significance of "significance":**

**When we reject the null hypothesis $\mathrm{H}_0$, we say that the result is *statistically significant*.**

**Discussion points:**

- **always report the p-value**
- **keep in mind that $\alpha = 0.05$ is arbitrary**
- **significance does not always mean importance**
- **p-values are related to sample size**

## More on stat significance vs practical importance:

**Example: Spring Birthday Confers Height Advantage - Yahoo Health News, Feb 18/98**

**In an Austrian study of 507,125 military recruits, it was found that the average height of those born in the spring was $1/4$ inch more than those born in the fall.**

# Lecture 30

## Errors in testing:

| | **Parameter Space** | |
|---|---|---|
| **Decision** | $\mathrm{H}_0$ **true** | $\mathrm{H}_1$ **true** |
| **Reject** $\mathrm{H}_0$ | **Type I Error** | |
| **Do not reject** $\mathrm{H}_0$ | | **Type II Error** |

## Discussion questions:

- **what is a good test?**
- **can we have a perfect test?**

**Example: We examine Type I error and Type II error in the earlier example where a defendent is accused of a crime in a court of law.**

## Probabilities associated with errors in testing:

| | **Parameter Space** | |
|---|---|---|
| **Decision** | $\mathrm{H}_0$ **true** | $\mathrm{H}_1$ **true** |
| **Reject** $\mathrm{H}_0$ | $\alpha$ | $1-\beta$ |
| **Do not reject** $\mathrm{H}_0$ | | $\beta$ |

## Discussion points:

- $\alpha$ **is the *significance level* of a test**
- **we typically fix** $\alpha$
- $1-\beta$ **is referred to as the *power* of a test**
- **we want the power to be large**
- $\alpha$, $\beta$ **are test properties; indpt of data**
- **note that in our examples,** $\mathrm{H}_0$ **is *simple***
- **note that in our examples,** $\mathrm{H}_1$ **is *composite***

**Example: We return to the one sample problem where $X_1, \ldots, X_n$ are iid, $\sigma = 1.8$, $\alpha = 0.05$ and $n = 100$. We are interested in testing $\mathrm{H}_0 : \mu = 3$ versus $\mathrm{H}_1 : \mu > 3$.**

**(a) Find the *critical region* (rejection region).**

**(b) Calculate the power at $\mu = 3.2$.**

**(c) Calculate the power at $\mu = 3.5$.**

**(d) What happens in (b) when $n = 100 \to 400$?**

**I have mentioned previously that statistical practice relies heavily on computation. Here is a general simulation procedure that can be used to approximate power. Suppose that you have data $X_1, \ldots, X_n$ and are testing $\mathrm{H}_0$ versus $\mathrm{H}_1$. Suppose that the critical region is $\{\underline{X} : \; Q(\underline{X}) \geq a\}$ and leads to an intractable power expression**

$$\begin{aligned} \text{Power} &= \mathrm{P}(\textbf{reject } \mathrm{H}_0 \mid \mathrm{H}_1 \textbf{ true}) \\ &= \mathrm{P}(Q(\underline{X}) \geq a \mid \mathrm{H}_1 \textbf{ true}) \end{aligned}$$

**Step 0: set counter = 0**

**Step 1: generate $x_1, \ldots, x_n$ under $\mathrm{H}_1$**

**Step 2: if $Q(\underline{x}) \geq a$, increase the counter by 1**

**Repeat Steps 1 and 2 M times and approximate**

$$\text{Power} \approx \frac{\text{counter}}{M}$$

# Lecture 31

**Example: Consider** $X \sim \mathrm{Bin}(500, p)$ **where we test** $\mathrm{H}_0 : p = 0.7$ **versus** $\mathrm{H}_1 : p \neq 0.7$ **at level** $\alpha = 0.01$**.**

**(a) Find the critical region of the test.**

**(b) Calculate the power at** $p = 0.6$**.**

**Example: In a two sample test of** $\mathrm{H}_0 : \mu_1 - \mu_2 = 3$ **versus** $\mathrm{H}_1 : \mu_1 - \mu_2 > 3$**, suppose that the data are normal,** $m = n$ **and** $\sigma_1^2 = \sigma_2^2 = 84.0$**. Can we choose** $m$ **such that the test has level** $\alpha = 0.01$ **and** $\beta = 0.05$ **at** $\mu_1 - \mu_2 = 5.0$**? This question concerns *experimental design.***

**In two sample problems, we can relax the normality assumption in the case of large samples.**

**Given $X_1, \ldots, X_m$ iid independent of $Y_1, \ldots Y_n$ iid with $m$ and $n$ large (ie. $m, n \geq 30$), then the following statistic can be used for testing and the construction of confidence intervals.**

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/m + s_2^2/n}} \sim \text{Normal}(0, 1)$$

**where $\mu_1$ and $\mu_2$ are the respective means, and $s_1$ and $s_2$ are the respective sample std devs.**

**Example: A college interviewed 1296 students wrt summer incomes. Based on the results in the following table, test whether there is a difference in earnings between male and female students.**

| Students | $n$ | $\bar{X}$ | $s$ |
|---|---|---|---|
| male | 675 | \$1884.52 | \$1368.37 |
| female | 621 | \$1360.39 | \$1037.46 |

**Example: The test scores of first year students admitted to college directly from high school historically exceed the test scores of first year students with working experience by 10%. A recent sample of 50 first year students admitted directly from high school has an average test score of 74.1% with std dev 3.8%. An indpt sample of 50 first year students with working experience yields an average test score of 66.5% with std dev 4.1%. Test whether a change has occurred.**

# Lecture 32

**We consider another variation to the two sample problem. This time, the data are again normal. Realistically, $\sigma_1$ and $\sigma_2$ are unknown but we need to make the additional assumption $\sigma_1 = \sigma_2$.**

**Given $X_1, \ldots, X_m$ iid $\mathrm{Normal}(\mu_1, \sigma_1^2)$ independent of $Y_1, \ldots Y_n$ iid $\mathrm{Normal}(\mu_2, \sigma_2^2)$ with $\sigma_1 = \sigma_2$, then the following statistic can be used for testing and the construction of confidence intervals.**

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}} \sim t_{m+n-2}$$

**where $s_1$ and $s_2$ are the respective sample std devs.**

**Example: The Chapin Social Insight Test gave the following scores. Assuming normal data, test whether the mean score of males exceeds the mean score of females.**

| Group | $n$ | $\bar{X}$ | $s$ |
|---|---|---|---|
| males | 18 | 25.34 | 13.36 |
| females | 23 | 24.94 | 14.39 |

## Example cont'd: Obtain a 95% CI for $\mu_1 - \mu_2$.

**There are actually lots of testing methodologies corresponding to different data scenarios. We will study one more situation (a common one involving paired data) but keep in mind that the principles that we have studied carry over to more complex situations.**

**Suppose in the paired data situation, we have $X_1, \ldots, X_n$ iid arising from a population with mean $\mu_1$, and $Y_1, \ldots, Y_n$ iid arising from a population with mean $\mu_2$. Furthermore, assume that the data are paired such that $X_i$ corresponds to $Y_i$. This natural pairing implies that there is a dependence between $X_i$ and $Y_i$.**

**To carry out inference (testing and the construction of CI's), we define a new random variable, the difference $D_i = X_i - Y_i$. Our interest concerns the unknown parameter**

$$\begin{aligned} \mathrm{E}(D_i) &= \mathrm{E}(X_i - Y_i) \\ &= \mathrm{E}(X_i) - \mathrm{E}(Y_i) \\ &= \mu_1 - \mu_2. \end{aligned}$$

**Our analysis proceeds as in the single sample case based on the data $D_1, \ldots, D_n$.**

**Example: Suppose scores measuring jitteriness are normally distributed . We believe that scores increase after drinking coffee. Let $X_i$ be the before drinking coffee score and let $Y_i$ be the the after drinking coffee score for the $i$-th individual. Based on $\alpha = 0.01$, test the hypothesis.**

| $x_i$ | $y_i$ | $d_i$ |
|---|---|---|
| 50 | 56 | |
| 60 | 70 | |
| 55 | 60 | |
| 72 | 70 | |
| 85 | 82 | |
| 78 | 84 | |
| 65 | 68 | |
| 90 | 88 | |

**Example cont'd: Obtain a 95% CI for the mean difference in jitteriness scores.**

**Example cont'd: Suppose we have the same data but the experiment involves 16 people where 8 people were measured without having coffee and 8 other people where measured after drinking coffee. How does the analysis differ?**

**Example cont'd: Suppose now that the 16 people involve 8 pairs of twins such that $X_i$ and $Y_i$ are twins. How should the analysis proceed?**

**Example cont'd: Assume the same conditions as above but the data are no longer normal. How should the analysis proceed?**

**Pairing is a special case of *blocking* (read in text). Blocking attempts to reduce variation by grouping data that are similar, and this hopefully leads to *more sensitive* tests (ie. tests that reject $\mathrm{H}_0$ more often when $\mathrm{H}_0$ is false).**

**Example: To illustrate the above, consider five before and after measurements involving a drug where there are big differences in responses between people but there is small variation in the $D_i$'s. Assuming normal data, we carry out a paired analysis and a non-paired analysis.**

| $x_i$ | $y_i$ | $d_i$ |
|---|---|---|
| 25 | 29 | $-4$ |
| 46 | 50 | $-4$ |
| 30 | 33 | $-3$ |
| 75 | 78 | $-3$ |
| 19 | 25 | $-6$ |

# Two Sample Testing - Summary

## Assume $X_1, \ldots, X_m$ iid with mean $\mu_1$ and std dev $\sigma_1$, and $Y_1, \ldots, Y_n$ iid with mean $\mu_2$ and std dev $\sigma_2$.

| Data | **Test Statistic** | **Comments** |
|---|---|---|
| **paired data**, $m = n$ | **take** $D_i = X_i - Y_i$ **and refer to single sample case** | |
| **non-paired**, $m, n$ **large** | $\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{N}(0, 1)$ | **replace** $\sigma_i$**'s with** $s_i$**'s if** $\sigma_i$**'s unknown** |
| **non-paired**, $m, n$ **not large, data normal,** $\sigma_i$**'s known** | $\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{N}(0, 1)$ | unrealistic |
| **non-paired**, $m, n$ **not large, data normal,** $\sigma_1 \approx \sigma_2$ **but unknown** | $\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s_p^2}} \sim t_{m+n-2}$ | $s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$ |
| **binomial data**, $m, n$ **large,** $p_1, p_2$ **moderate** | $\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/m + p_2(1-p_2)/n}} \sim \mathrm{normal}(0, 1)$ | **replace** $p$**'s with estimates in denominator** |