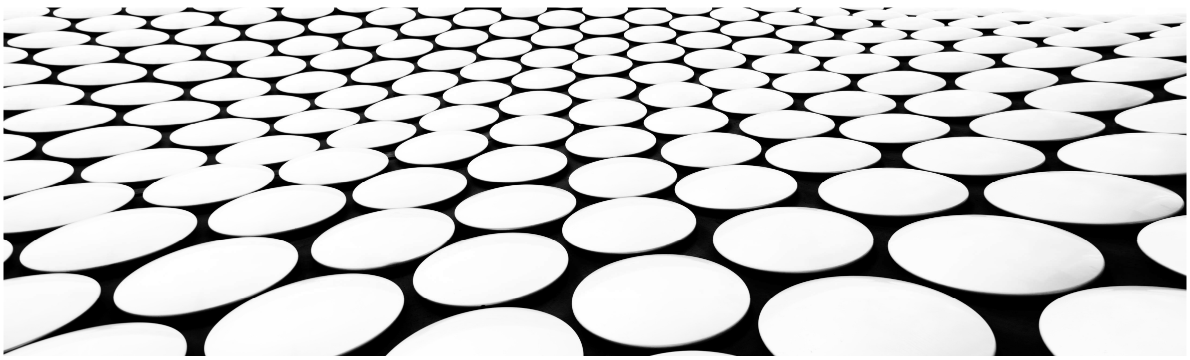
CSE353 – MACHINE LEARNING
MATHEMATICAL PRELIMINARIES

PRAVIN PAWAR, SUNY KOREA

SOME OF THE SLIDES ARE USED WITH PERMISSION FROM:

KRISTIN L. SAINANI, ASSOCIATE PROFESSOR, HEALTH AND RESEARCH POLICY, STANFORD UNIVERSITY.

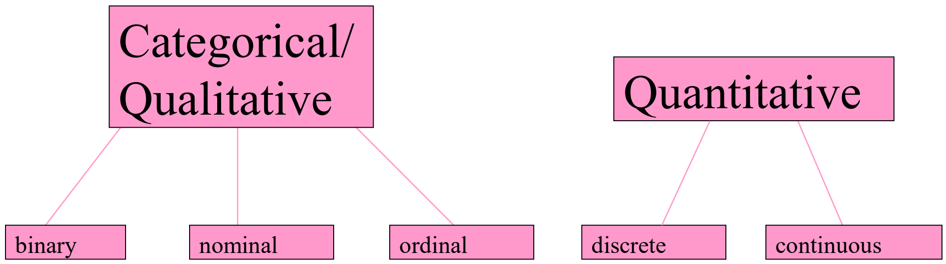


1

DESCRIPTIVE STATISTICS

2

TYPES OF VARIABLES: OVERVIEW
DISCUSS AND GIVE EXAMPLE OF EACH



2 categories +

more categories +

order matters +

numerical +

uninterrupted

3

LOOKING AT DATA

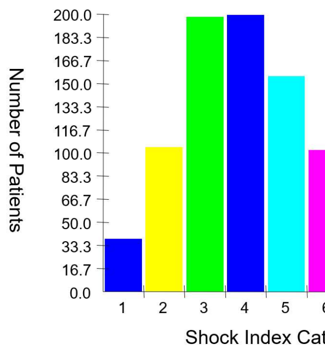
- ✓ How are the data distributed?
  - Where is the center?
  - What is the range?
  - What's the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?
- ✓ Are there "outliers"?
- ✓ Are there data points that don't make sense?
- ✓ 90% information is contained in the graph

4

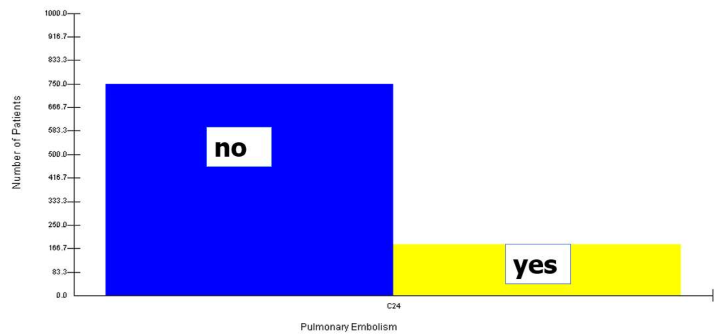
FREQUENCY PLOTS

Categorical variables

- Bar Chart
  - Used for categorical variables to show frequency or proportion in each category.
  - Translate the data from frequency tables into a pictorial representation



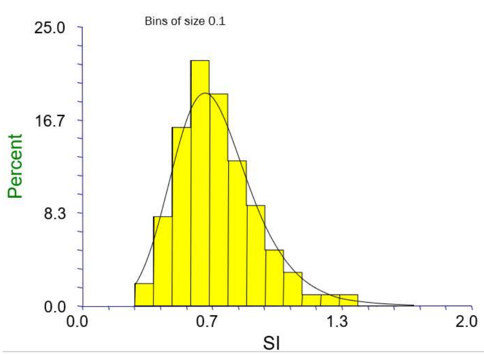
Much easier to extract information from a bar chart than from a table!

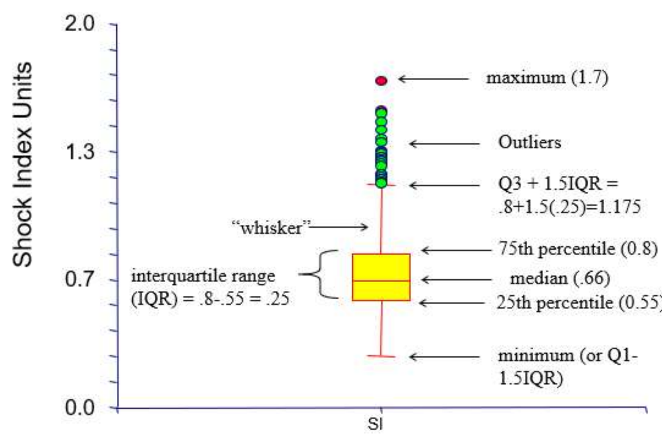


5

BOX PLOT AND HISTOGRAMS: FOR CONTINUOUS VARIABLES

Continuous variables
- Box Plot
- Histogram
- To show the distribution (shape, center, range, variation) of continuous variables.

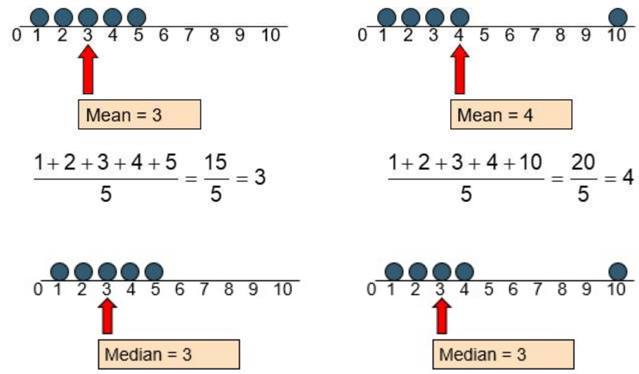




6

MEASURES OF CENTRAL TENDENCY

- Mean
  - The average; the balancing point
  - The mean is affected by extreme values/outliers
- Median
  - The exact middle value
  - In case of even observations, take middle two and average them
  - The median is not affected by extreme values/outliers
- Mode
  - The value that appears most frequently

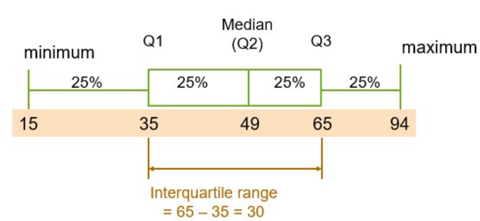


7

MEASURES OF VARIATION/DISPERSION

- Range
  - Difference between the largest and the smallest observations.
- Percentiles/quartiles
  - The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
  - Q_2 is the same as the median (50% are smaller, 50% are larger)
  - Only 25% of the observations are greater than the third quartile
- Interquartile range
  - Interquartile range = 3rd quartile – 1st quartile = $Q_3 - Q_1$
- Standard deviation
  - Most commonly used measure of variation
  - Shows variation about the mean
  - Has the same units as the original data
- Variance
  - Average (roughly) of squared deviations of values from the mean

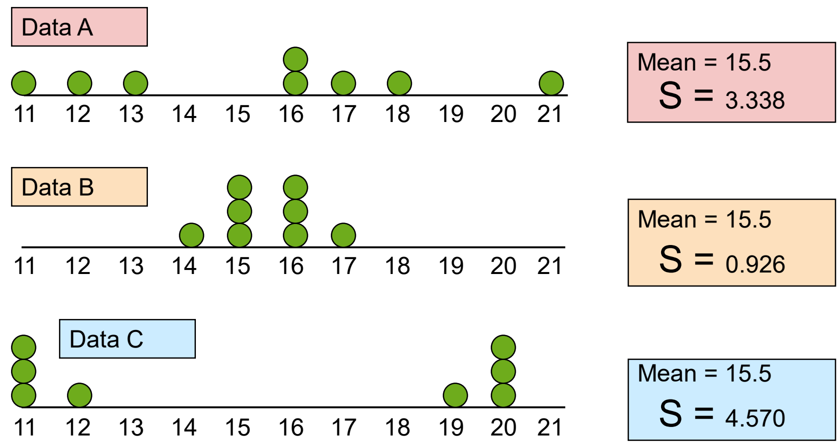
$$S = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{X})^2}{n-1}}$$



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

8

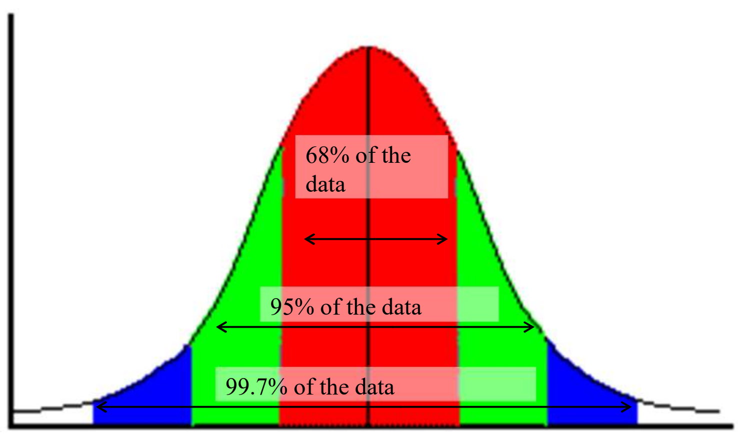
COMPARING STANDARD DEVIATIONS



9

THE BEAUTY OF THE NORMAL CURVE - 68-95-99.7 RULE

- No matter what μ and σ are, the area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%
- The area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%; and
- The area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%.
- Almost all values fall within 3 standard deviations.



10

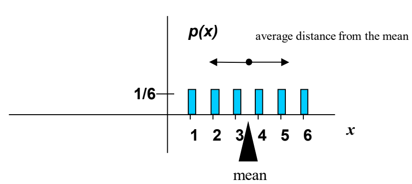
SUMMARY OF SYMBOLS

- S^2= Sample variance
- S = Sample standard dev
- σ^2 = Population (true or theoretical) variance
- σ = Population standard dev.
- $\overline{X}$ = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)

11

WHAT'S THE VARIANCE AND STANDARD DEVIATION OF THE ROLL OF A DIE?

x	*p(x)*
1	*p(x=1)=1/6*
2	*p(x=2)=1/6*
3	*p(x=3)=1/6*
4	*p(x=4)=1/6*
5	*p(x=5)=1/6*
6	*p(x=6)=1/6*
	1.0



$$E(x) = \sum_{\text{all } x} x_i\, p(x_i) = (1)(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + 4(\frac{1}{6}) + 5(\frac{1}{6}) + 6(\frac{1}{6}) = \frac{21}{6} = 3.5$$

$$E(x^2) = \sum_{\text{all } x} x_i^{\,2}\, p(x_i) = (1)(\frac{1}{6}) + 4(\frac{1}{6}) + 9(\frac{1}{6}) + 16(\frac{1}{6}) + 25(\frac{1}{6}) + 36(\frac{1}{6}) = 15.17$$

$$\sigma_x^2 = Var(x) = E(x^2) - [E(x)]^2 = 15.17 - 3.5^2 = 2.92$$

$$\sigma_x = \sqrt{2.92} = 1.71$$

12

VARIANCE PROPERTIES

If c= a constant number (i.e., not a variable) and *X* and *Y* are random variables, then

- Var(c) = 0
- Var (c+*X*)= Var(*X*)
- Var(c*X*)= c^2Var(*X*)
- Var(*X*+*Y*)= Var(*X*) + Var(*Y*)

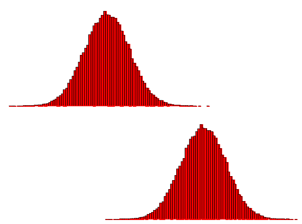
ONLY IF X and Y are independent!!!!

- {Var(*X*+*Y*)= Var(*X*) + Var(*Y*)+2Cov(X,Y)

IF X and Y are not independent}

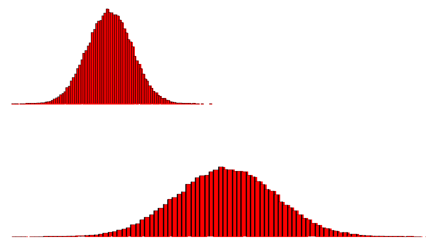
Var (c+*X*)= Var(*X*)

Adding a constant to every instance of a random variable doesn't change the variability. It just shifts the whole distribution by c. If everybody grew 5 inches suddenly, the variability in the population would still be the same.



Var(cX)= c^2Var(X)

Multiplying each instance of the random variable by c makes it c-times as wide of a distribution, which corresponds to c^2 as much variance (deviation squared). For example, if everyone suddenly became twice as tall, there'd be twice the deviation and 4 times the variance in heights in the population.



13

PRACTICE PROBLEM

Find the variance and standard deviation for the number of ships to arrive at the harbor (recall that the mean is 11.3).

x	10	11	12	13	14
P(x)	**.4**	**.2**	**.2**	**.1**	**.1**

14

VARIANCE AND STD DEV EXAMPLE

Find the variance and standard deviation for the number of ships to arrive at the harbor (recall that the mean is 11.3).

x^2	100	121	144	169	196
P(x)	.4	.2	.2	.1	.1

$$E(x^2) = \sum_{i=1}^{5} x_i^{\ 2}\, p(x_i) = (100)(.4) + (121)(.2) + 144(.2) + 169(.1) + 196(.1) = 129.5$$

$$Var(x) = E(x^2) - [E(x)]^2 = 129.5 - 11.3^2 = 1.81$$

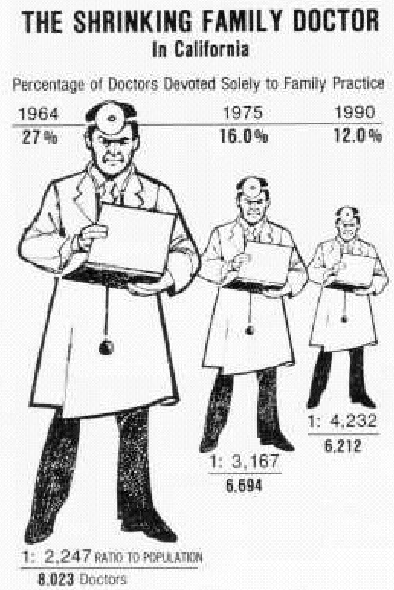
$$stddev(x) = \sqrt{1.81} = 1.35$$

Interpretation: On an average day, we expect 11.3 ships to arrive in the harbor, plus or minus 1.35. This gives you a feel for what would be considered a usual day!

15

EXAMPLES OF BAD GRAPHICS

16

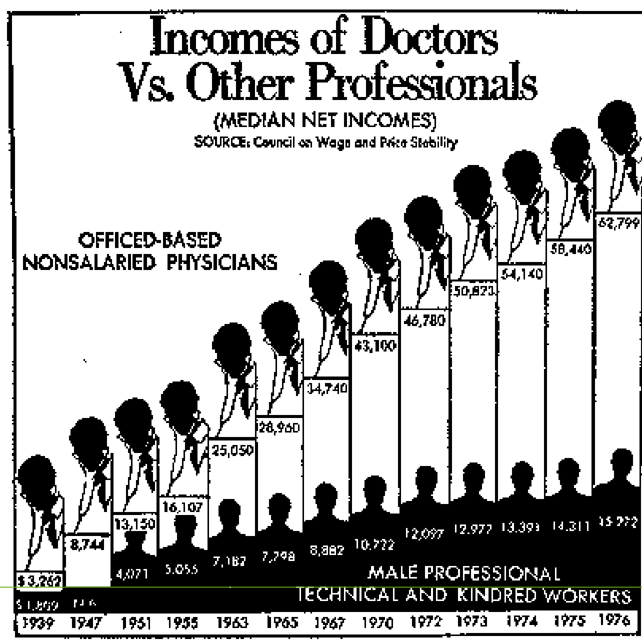


What's wrong with this graph?

Los Angeles Times, August 5, 1979, p. 3.

from: ER Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983, p.69

17

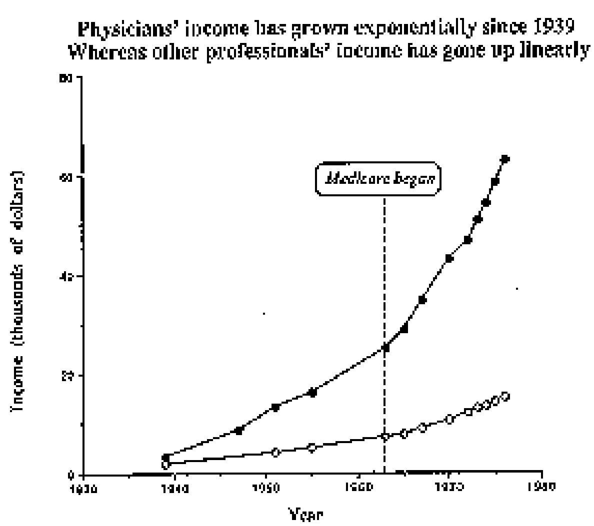


Notice the X-axis

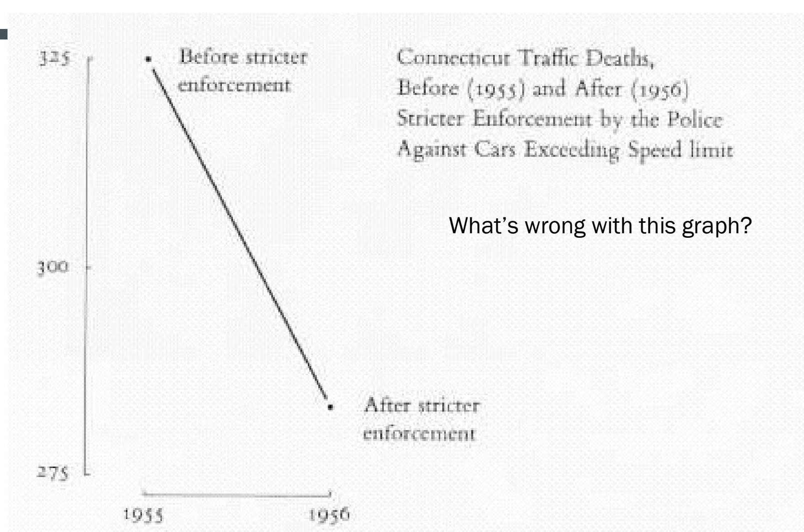
From: Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot
Wainer, H. 1997, p.29.

18

Correctly scaled X-axis…

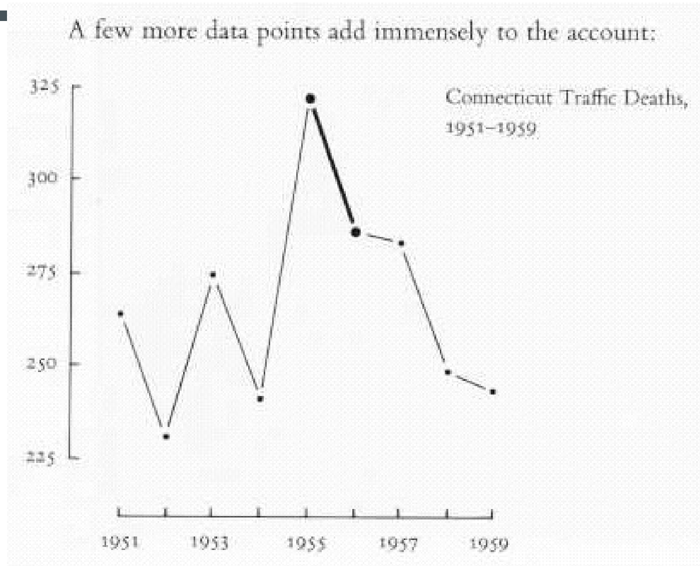


19



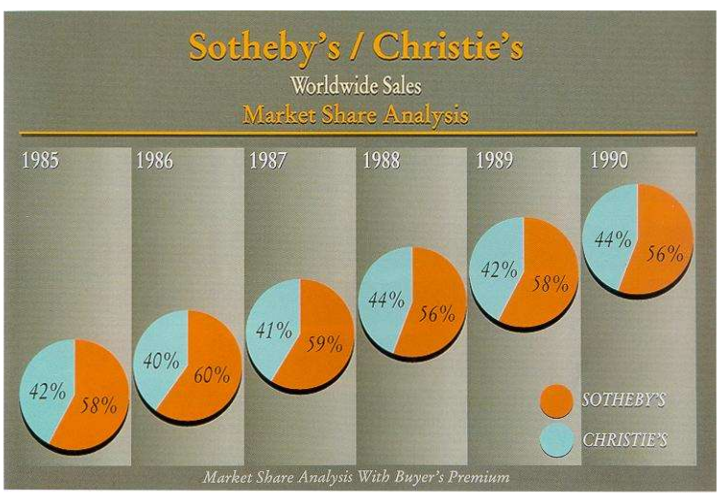
from: ER Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 1983, p.74

20



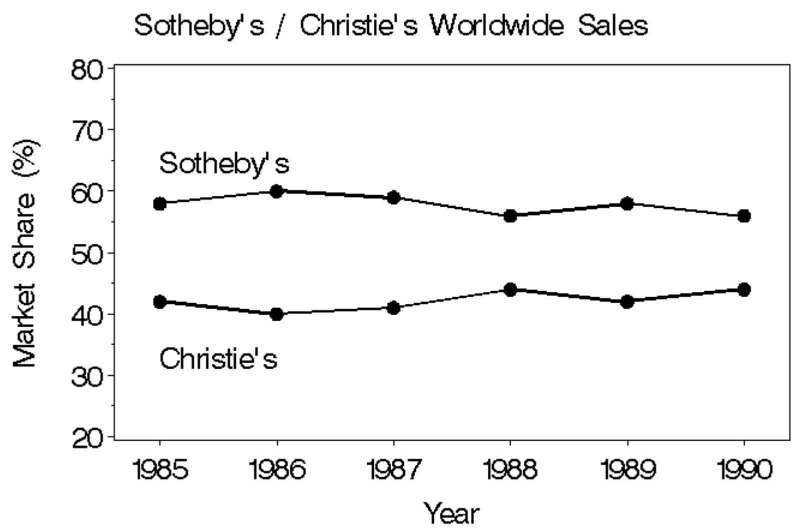
21

What's the message here?



Diagraphics II, 1994

22



Diagraphics II, 1994

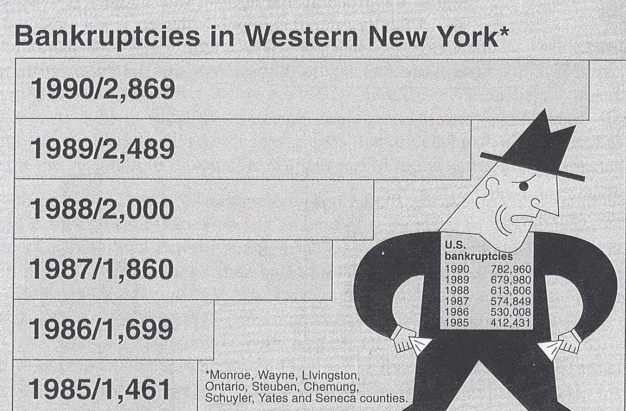
23

MORE PEOPLE FILE: BLAME RECESSION AND EASY CREDIT

By Janet Lively

Last Wednesday, a substance abuse counselor, a single mother on disability and the owner of a foreign automotive repair business gave up any hope of paying their bills.

They filed for bankruptcy, joining the more than 800 people and businesses who have asked for relief this year from the Western New York District of the U.S. Bankruptcy Court in Rochester. If filings continue at the same rate, 1991 will easily be another record year for the court.



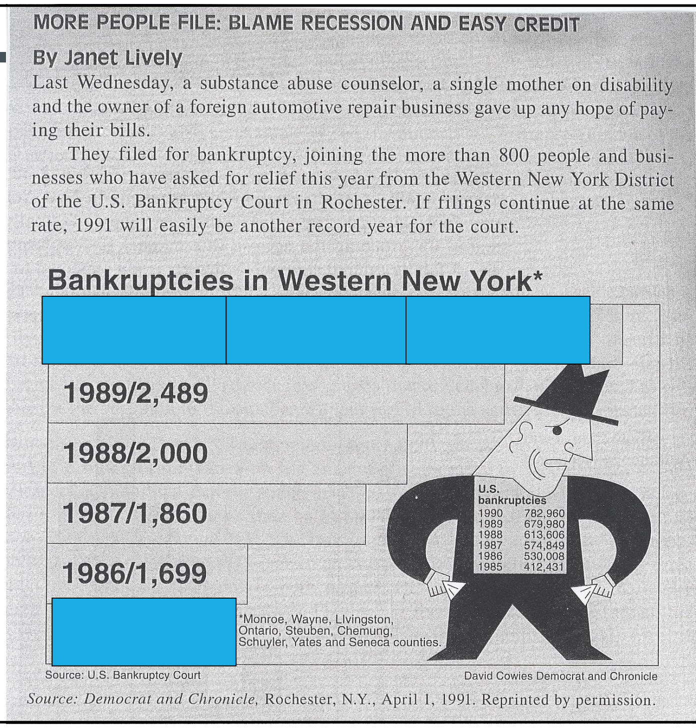
Source: U.S. Bankruptcy Court

David Cowies Democrat and Chronicle

Source: Democrat and Chronicle, Rochester, N.Y., April 1, 1991. Reprinted by permission.

From: Johnson R. *Just the Essentials of Statistics*. Duxbury Press, 1995.

24


MORE PEOPLE FILE: BLAME RECESSION AND EASY CREDIT

By Janet Lively

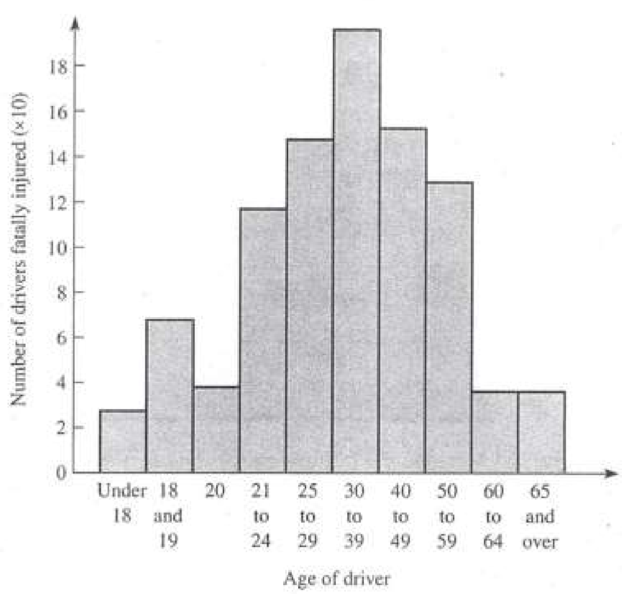
Last Wednesday, a substance abuse counselor, a single mother on disability and the owner of a foreign automotive repair business gave up any hope of paying their bills.

They filed for bankruptcy, joining the more than 800 people and businesses who have asked for relief this year from the Western New York District of the U.S. Bankruptcy Court in Rochester. If filings continue at the same rate, 1991 will easily be another record year for the court.

Source: Democrat and Chronicle, Rochester, N.Y., April 1, 1991. Reprinted by permission.

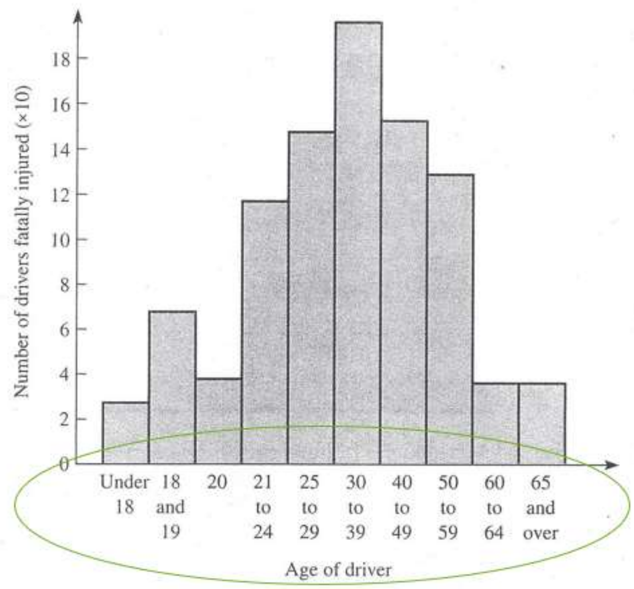
From: Johnson R. *Just the Essentials of Statistics*. Duxbury Press, 1995.

25



From: Johnson R. *Just the Essentials of Statistics*. Duxbury Press, 1995.

26



From: Johnson R. *Just the Essentials of Statistics*. Duxbury Press, 1995.

27

PROBABILITY DISTRIBUTIONS

- Random variables
- Probability functions
- Expected value
- Covariance

28

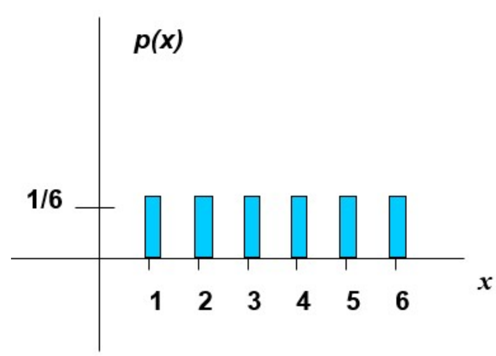
RANDOM VARIABLE

- Roughly, probability is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)
- A random variable *x* takes on a defined set of values with different probabilities.
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is also a random variable.
- Discrete random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- Continuous random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

29

PROBABILITY FUNCTIONS

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.



$$\sum_{\text{all } x} P(x) = 1$$

30

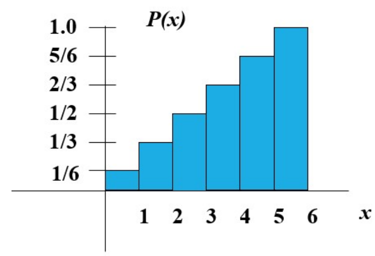
PROBABILITY MASS FUNCTION (PMF)

x	$p(x)$
1	$p(x=1)$=1/6
2	$p(x=2)$=1/6
3	$p(x=3)$=1/6
4	$p(x=4)$=1/6
5	$p(x=5)$=1/6
6	$p(x=6)$=1/6

1.0

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

x	$P(x \le A)$
1	$P(x \le 1)$=1/6
2	$P(x \le 2)$=2/6
3	$P(x \le 3)$=3/6
4	$P(x \le 4)$=4/6
5	$P(x \le 5)$=5/6
6	$P(x \le 6)$=6/6



31

EXAMPLES

1. What's the probability that you roll a 3 or less?
$P(x \le 3)$=1/2

2. What's the probability that you roll a 5 or higher?
$P(x \ge 5)$ = 1 – $P(x \le 4)$ = 1-2/3 = 1/3

Which of the following are probability functions?
Hint: The sum of all probabilities is 1 and there is no negative probability.

a. $f(x)=.25$ for x=9,10,11,12 (YES)

b. $f(x)= (3-x)/2$ for x=1,2,3,4 (NO)

c. $f(x)= (x^2+x+1)/25$ for x=0,1,2,3 (NO)

32

PRACTICE PROBLEM:

- The number of ships to arrive at a harbor on any given day is a random variable represented by *x*. The probability distribution for *x* is:

x	**10**	**11**	**12**	**13**	**14**
P(x)	**.4**	**.2**	**.2**	**.1**	**.1**

Find the probability that on a given day:

a. exactly 14 ships arrive $p(x=14) = .1$

b. At least 12 ships arrive $p(x \geq 12) = (.2 + .1 + .1) = .4$

c. At most 11 ships arrive $p(x \leq 11) = (.4 + .2) = .6$

33

PRACTICE PROBLEM:

You are lecturing to a group of 1000 students. You ask them to each randomly pick an integer between 1 and 10. Assuming, their picks are truly random:

- What's your best guess for how many students picked the number 9?

  Since $p(x=9)$ = 1/10, we'd expect about 1/10th of the 1000 students to pick 9. 100 students.

- What percentage of the students would you expect picked a number less than or equal to 6?

  Since $p(x\leq 6)$ = 1/10 + 1/10 + 1/10 + 1/10 + 1/10 + 1/10 =.6 (60%)

34

CONTINUOUS CASE

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
- The probabilities associated with continuous functions are just areas under the curve (integrals!).
- Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting a math SAT score between 700 and 800 is 2%).

35

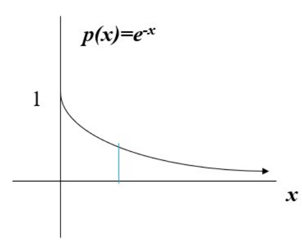
CONTINUOUS CASE EXAMPLE

$$f(x) = e^{-x}$$

- For example, recall the negative exponential function (in probability, this is called an "exponential distribution"):
- This function integrates to 1.
- e is approximately equal to 2.71828.

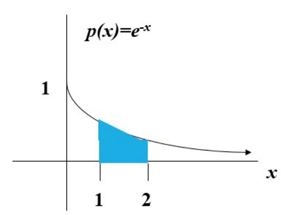
$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$

- The probability that x is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of x.



- For example, the probability of x falling within 1 to 2:

$$\boxed{P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23}$$



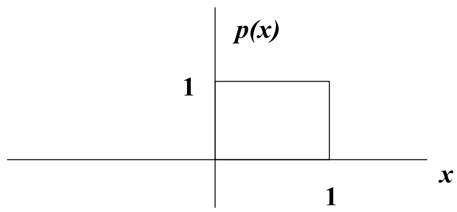
$$P(x \le 2) = 1 - e^{-2} = 1 - .135 = .865$$

36

EXAMPLE 2: UNIFORM DISTRIBUTION

- The uniform distribution: all values are equally likely
- The uniform distribution:

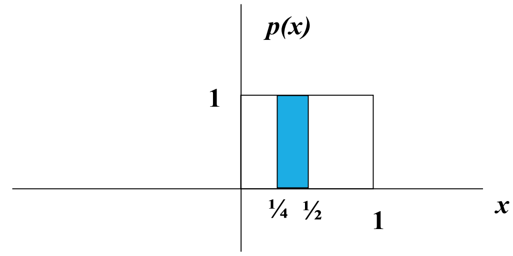
$f(x) = 1$, for $1 \geq x \geq 0$



- We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

What's the probability that *x* is between ¼ and ½?



P(½ ≥*x*≥ ¼)= ¼

37

PRACTICE PROBLEM

- Suppose that survival drops off rapidly in the year following diagnosis of a certain type of advanced cancer. Suppose that the length of survival (or time-to-death) is a random variable that approximately follows an exponential distribution with parameter 2 (makes it a steeper drop off):

$$\text{probability function}: p(x = T) = 2e^{-2T}$$

$$[note: \int_0^{+\infty} 2e^{-2x} = -e^{-2x} \Big|_0^{+\infty} = 0 + 1 = 1]$$

- What's the probability that a person who is diagnosed with this illness survives a year?

The probability of dying within 1 year can be calculated using the cumulative distribution function:

$$P(x \leq T) = -e^{-2x} \Big|_0^T = 1 - e^{-2(T)}$$

$$1 - (1 - e^{-2(1)}) = .135$$

38

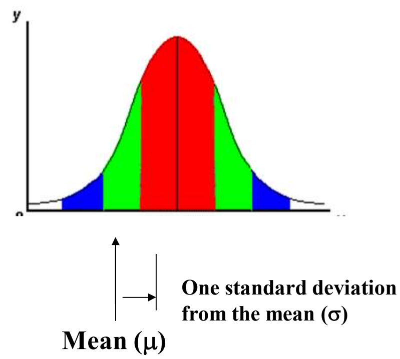
EXPECTED VALUE AND VARIANCE

- All probability distributions are characterized by an expected value and a variance (standard deviation squared).
- If we understand the underlying probability function of a certain phenomenon, then we can make informed decisions based on how we expect *x* to behave on-average over the long-run…(so called "frequentist" theory of probability).
- Expected value is just the weighted average or mean (μ) of random variable *x*. Imagine placing the masses *p*(*x*) at the points *X* on a beam; the balance point of the beam is the expected value of *x*.
- Recall the following probability distribution of ship arrivals:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

$$\sum_{i=1}^{5} x_i p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Bell-curve (normal) distribution



39

EXPECTED VALUE, FORMALLY

Discrete case:

$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

General Formula:

Outcome	X_1	X_2	X_3	X_4	…	X_n
Probability	P_1	P_2	P_3	P_4	…	P_n

Expected Value = $P_1X_1 + P_2X_2 + P_3X_3 + P_4X_4 + \dots + P_nX_n$

Example: A single fair six-sided die is rolled.

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

Expected Value = $1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

E(X) = μ
Can be used interchangeably.

Expected value is an extremely useful concept for good decision-making!

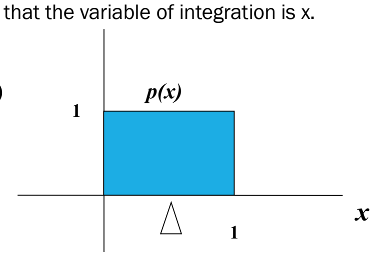
Continuous case:

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

The symbol dx, called the differential of the variable x, indicates that the variable of integration is x.

Continuous case (uniform distribution)

$$E(X) = \int_0^1 x(1)dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$



40

EXAMPLE: THE LOTTERY

- The Lottery (also known as a tax on people who are bad at math…)
- A certain lottery works by picking 6 numbers from 1 to 49. It costs $1.00 to play the lottery, and if you win, you win $2 million after taxes.
- *If you play the lottery once, what are your expected winnings or losses?*

Calculate the probability of winning in 1 try:

$$\frac{1}{\binom{49}{6}} = \frac{1}{\frac{49!}{43!6!}} = \frac{1}{13,983,816} = 7.2 \times 10^{-8}$$

"49 choose 6"

Out of 49 numbers, this is the number of distinct combinations of 6.

If you play the lottery every week for 10 years, what are your expected winnings or losses?

520 x (-.86) = -$447.20

The probability function (note, sums to 1.0):

x$	*p(x)*
-1	**.999999928**
+ 2 million	***7.2 x 10⁻⁸***

Expected Value

E(X) = P(win)*$2,000,000 + P(lose)*-$1.00
= $2.0 \times 10^6 * 7.2 \times 10^{-8} + .999999928\,(-1) = .144 - .999999928 = -\$.86$

Negative expected value is never good!
You shouldn't play if you expect to lose money!

41

GAMBLING (OR HOW CASINOS CAN AFFORD TO GIVE SO MANY FREE DRINKS...)

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet $1 that an odd number comes up, you win or lose $1 according to whether or not that event occurs. If random variable X denotes your net gain, X=1 with probability 18/38 and X= -1 with probability 20/38.

E(X) = 1(18/38) – 1 (20/38) = -$.053

On average, the casino wins (and the player loses) 5 cents per game.

The casino rakes in even more if the stakes are higher:

E(X) = 10(18/38) – 10 (20/38) = -$.53

If the cost is $10 per game, the casino wins an average of 53 cents per game. If 10,000 games are played in a night, that's a cool $5300.

42

EXPECTED VALUE OF A COIN TOSS

You toss a coin 100 times. What's the expected number of heads? What's the variance of the number of heads?

Intuitively, we'd probably all agree that we expect around 50 heads, right?

Another way to show this→

Think of tossing 1 coin. E(X=number of heads) = (1) P(heads) + (0)P(tails)

∴ E(X=number of heads) = 1(.5) + 0 = .5

If we do this 100 times, we're looking for the sum of 100 tosses, where we assign 1 for a heads and 0 for a tails.

$E(X_1 + X_2 + X_3 + X_4 + X_5 \ldots + X_{100}) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \ldots + E(X_{100}) =$

$100\, E(X_1) = 50$

43

VARIANCE IN 100 COIN TOSS

What's the variability, though? More tricky. But, again, we could do this for 1 coin and then use our rules of variance.

Think of tossing 1 coin.

$E(X^2$=number of heads squared$) = 1^2$ P(heads) + 0^2 P(tails)

∴ $E(X^2) = 1(.5) + 0 = .5$

$Var(X) = .5 - .5^2 = .5 - .25 = .25$

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

Then, using our rule: Var(X+Y)= Var(X) + Var(Y) (coin tosses are independent!)

$Var(X_1 + X_2 + X_3 + X_4 + X_5 \ldots + X_{100}) = Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \ldots + Var(X_{100}) =$

$100\, Var(X_1) = 100\,(.25) = 25$

SD(X)=5

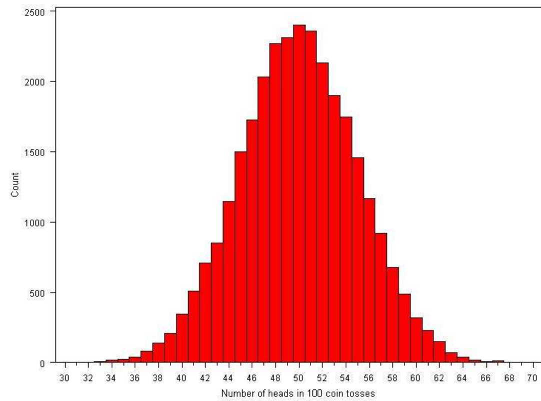
The variance of X is equal to the mean of the square of X minus the square of the mean of X.

Interpretation: When we toss a coin 100 times, we expect to get 50 heads plus or minus 5.

44

OR USE COMPUTER SIMULATION…

- Flip coins virtually!
  - Flip a virtual coin 100 times; count the number of heads.
  - Repeat this over and over again a large number of times (we'll try 30,000 repeats!)
  - Plot the 30,000 results.



Mean = 50

Std. dev = 5

Follows a normal distribution

∴ 95% of the time, we get between 40 and 60 heads…

45

COVARIANCE: JOINT PROBABILITY

- The covariance measures the strength of the linear relationship between two variables
- The covariance: $E[(x-\mu_x)(y-\mu_y)]$

$$\sigma_{xy} = \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)\, P(x_i, y_i)$$

- The sample covariance:

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

- **Covariance** between two random variables:

cov(X,Y) > 0 ⟶ X and Y are positively correlated

cov(X,Y) < 0 ⟶ X and Y are inversely correlated

cov(X,Y) = 0 ⟶ X and Y are independent

46

BAYES' THEOREM

- Conditional probability: Conditional probability is denoted by $P(B=b|A=a)$. It is the probability of $B=b$, provided that $A=a$ has occurred.
- Multiplication rule: It is denoted by $P(A,B)=P(B|A)P(A)$. It says that the probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.
- Bayes' theorem is given by:

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

  - where A and B are two events and P(B) ≠ 0
  - P(A | B) is the conditional probability of event A occurring given that B is true.
  - P(B | A) is the conditional probability of event B occurring given that A is true.
  - P(A) and P(B) are the probabilities of A and B occurring independently of one another.

- Examples: https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

47

APPLICATIONS OF PROBABILITY

Assume that a doctor administers an AIDS test to a patient. This test is fairly accurate and it fails only with 1% probability if the patient is healthy but reporting him as diseased. Moreover, it never fails to detect HIV if the patient actually has it. We use D_1 to indicate the diagnosis (1 if positive and 0 if negative) and H to denote the HIV status (1 if positive and 0 if negative).

:Conditional probability of $P(D_1 \mid H)$.

Conditional probability	$H = 1$	$H = 0$
$P(D_1 = 1 \mid H)$	1	0.01
$P(D_1 = 0 \mid H)$	0	0.99

- What is the probability the patient has AIDS if the test comes back positive, i.e., $P(H=1|D_1=1)$?
- Assume that the population is quite healthy, e.g., $P(H=1)=0.0015$.

48

ANSWER

To apply Bayes' theorem, we need to determine

$$\begin{aligned}&P(D_1 = 1)\\=&P(D_1 = 1, H = 0) + P(D_1 = 1, H = 1)\\=&P(D_1 = 1 \mid H = 0)P(H = 0) + P(D_1 = 1 \mid H = 1)P(H = 1)\\=&0.011485.\end{aligned}$$

Thus, we get

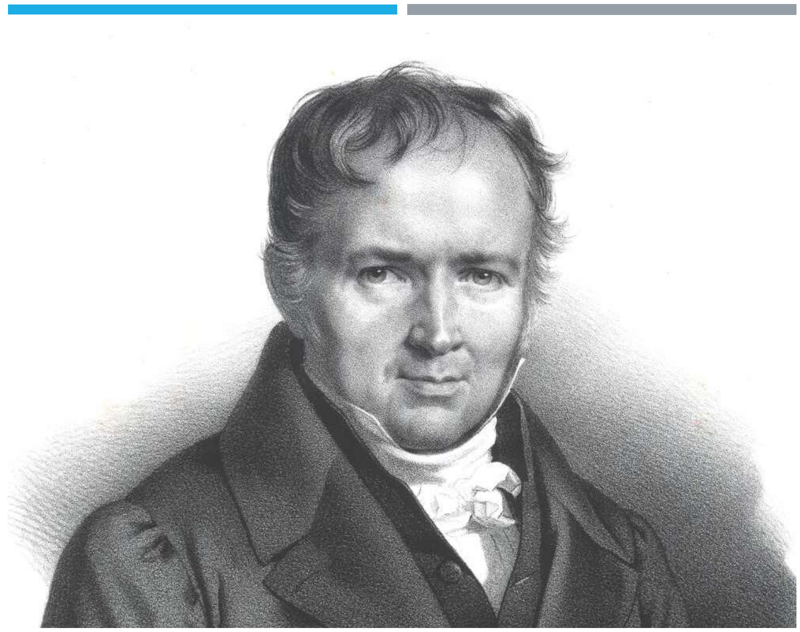
$$\begin{aligned}&P(H = 1 \mid D_1 = 1)\\=&\frac{P(D_1 = 1 \mid H = 1)P(H = 1)}{P(D_1 = 1)}.\\=&0.1306\end{aligned}$$

In other words, there is only a 13.06% chance that the patient actually has AIDS, despite using a very accurate test. As we can see, probability can be counterintuitive.

49

IMPORTANT DISCRETE DISTRIBUTIONS

- Binomial
  - Yes/no outcomes (dead/alive, treated/untreated, smoker/non-smoker, sick/well, etc.)
- Poisson
  - Counts (e.g., how many cases of disease in a given area)



50

BINOMIAL PROBABILITY DISTRIBUTION

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary random variable
  - e.g., head or tail in each toss of a coin; defective or not defective light bulb
  - Generally called "success" and "failure"
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

51

BINOMIAL EXAMPLE

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this exact arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) $=(1/2)^3 \times (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 \times (1/2)^3 \times (1/2)^1 = (1/2)^3 \times (1/2)^2$

In fact, $(1/2)^3 \times (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + \ldots\ldots$ for as many unique arrangements as there are—but how many are there??

52

Outcome	Probability
THHHT	$(1/2)^3$ *x* $(1/2)^2$
HHHTT	$(1/2)^3$ *x* $(1/2)^2$
TTHHH	$(1/2)^3$ *x* $(1/2)^2$
HTTHH	$(1/2)^3$ *x* $(1/2)^2$
HHTTH	$(1/2)^3$ *x* $(1/2)^2$
THTHH	$(1/2)^3$ *x* $(1/2)^2$
HTHTH	$(1/2)^3$ *x* $(1/2)^2$
HHTHT	$(1/2)^3$ *x* $(1/2)^2$
THHTH	$(1/2)^3$ *x* $(1/2)^2$
HTHHT	$(1/2)^3$ *x* $(1/2)^2$

10 arrangements *x* $(1/2)^3$ *x* $(1/2)^2$

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

${}_5C_3 = 5!/3!2! = 10$

The probability of each unique outcome (note: they are all equal)

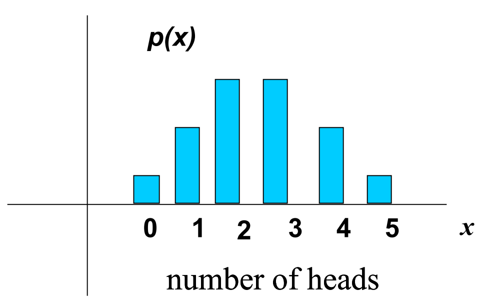
$\therefore$ **P(3 heads and 2 tails) =** ***x P(heads)^3 x P(tails)^2 =***

10 x (½)^5=31.25%

53

BINOMIAL DISTRIBUTION FUNCTION:

X= THE NUMBER OF HEADS TOSSED IN 5 COIN TOSSES



Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials, then the probability of exactly *X* "successes"=

$$\binom{n}{X} p^X (1-p)^{n-X}$$

n = number of trials

1-p = probability of failure

X = # successes out of *n* trials

p = probability of success

54

DEFINITIONS: BINOMIAL

- **Binomial:** Suppose that *n* independent experiments, or trials, are performed, where *n* is a fixed number, and that each experiment results in a "success" with probability *p* and a "failure" with probability *1-p*. The total number of successes, *X*, is a binomial random variable with parameters *n* and *p*.
- We write: ***X ~ Bin (n, p)*** *{reads: "X is distributed binomially with parameters n and p}*
- And the probability that $X=r$ (i.e., that there are exactly *r* successes) is:

$$P(X=r)=\binom{n}{r}p^r(1-p)^{n-r}$$

DEFINITIONS: BERNOULLI

- **Bernouilli trial:** If there is only 1 trial with probability of success p and probability of failure 1-p, this is called a Bernouilli distribution. (special case of the binomial with n=1)

Probability of success $P(X=1)=\binom{1}{1}p^1(1-p)^{1-1}=p$

Probability of failure $P(X=0)=\binom{1}{0}p^0(1-p)^{1-0}=1-p$

55

BINOMIAL DISTRIBUTION: EXAMPLE

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10}(.5)^{10}(.5)^{10}=.176$$

- If I toss a coin 20 times, what's the probability of getting of getting 2 or fewer heads?

$$\binom{20}{0}(.5)^0(.5)^{20}=\frac{20!}{20!0!}(.5)^{20}=9.5x10^{-7}+$$

$$\binom{20}{1}(.5)^1(.5)^{19}=\frac{20!}{19!1!}(.5)^{20}=20x9.5x10^{-7}=1.9x10^{-5}+$$

$$\binom{20}{2}(.5)^2(.5)^{18}=\frac{20!}{18!2!}(.5)^{20}=190x9.5x10^{-7}=1.8x10^{-4}$$

$$=1.8x10^{-4}$$

56

MULTINOMIAL DISTRIBUTION

The multinomial is a generalization of the binomial. It is used when there are more than 2 possible outcomes (for ordinal or nominal, rather than binary, random variables).

- Instead of partitioning *n* trials into 2 outcomes (yes with probability *p* / no with probability *1-p*), you are partitioning *n* trials into 3 or more outcomes (with probabilities: p_1, p_2, p_3,..)
- General formula for 3 outcomes:

$$P(D=x,R=y,G=z)=\frac{n!}{x!\,y!\,z!}p_D^x p_R^y (1-p_D-p_R)^z$$

Specific Example: if you are randomly choosing 8 people from an audience that contains 50% democrats, 30% republicans, and 20% green party, what's the probability of choosing exactly 4 democrats, 3 republicans, and 1 green party member?

$$P(D=4,R=3,G=1)=\frac{8!}{4!\,3!\,1!}(.5)^4(.3)^3(.2)^1$$

57

POISSON DISTRIBUTION

- Poisson distribution is for counts—if events happen at a constant rate over time, the Poisson distribution gives the probability of X number of events occurring in time T.

POISSON MEAN AND VARIANCE

- Mean $\mu=\lambda$

For a Poisson random variable, the variance and mean are the same!

- Variance and Standard Deviation

$$\sigma^2=\lambda$$

$$\sigma=\sqrt{\lambda}$$

where λ = expected number of hits in a given time period

58

POISSON DISTRIBUTION EXAMPLE

The Poisson distribution models counts, such as the number of new cases of SARS that occur in women in New England next month.

The distribution tells you the probability of all possible numbers of new cases, from 0 to infinity.

If X= # of new cases next month and *X* ~ Poisson (λ), then the probability that *X=k* (a particular count) is:

$$p(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

For example, if new cases of West Nile Virus in New England are occurring at a rate of about 2 per month, then these are the probabilities that: 0,1, 2, 3, 4, 5, 6, to 1000 to 1 million to… cases will occur in New England in the next month:

X	P(X)
0	$\frac{2^0 e^{-2}}{0!}$ =.135
1	$\frac{2^1 e^{-2}}{1!}$ =.27
2	$\frac{2^2 e^{-2}}{2!}$ =.27
3	$\frac{2^3 e^{-2}}{3!}$ =.18
4	=.09
5	
…	…

59

MORE ON POISSON…

"Poisson Process" (rates)

Note that the Poisson parameter λ can be given as the mean number of events that occur in a defined time period OR, equivalently, λ can be given as a rate, such as λ=2/month (2 events per 1 month) that must be multiplied by *t*=time (called a "Poisson Process") →

X ~ Poisson (λ)

$$P(X=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

E(X) = λt
Var(X) = λt

1a. If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

X ~ Poisson (λ=2 calls/hour)

P(X≥1)=1 – P(X=0)

$$P(X=0)=\frac{(2*1.5)^0 e^{-2(1.5)}}{0!}\frac{(3)^0 e^{-3}}{0!}=e^{-3}=.05$$

∴P(X≥1)=1 – .05 = 95% chance

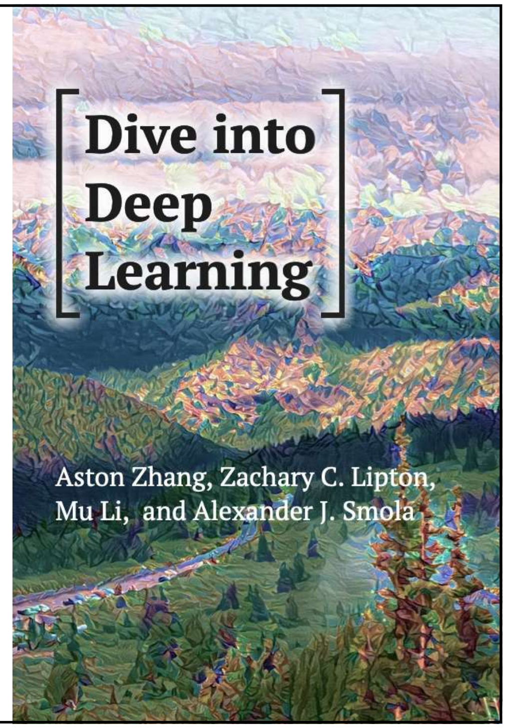
1b. How many phone calls do you expect to get during the movie?

E(X) = λt = 2(1.5) = 3

60

GOAL

- Data storage, manipulation and preprocessing are fundamental to machine learning.
- This lecture provides a rapid introduction to basic and frequently-used mathematics used in machine learning
- Matrix operations and their implementations
- Bit of calculus and probability
- For further understanding of all of the mathematical content, review Chapter 18 (Appendix - Mathematics for Deep Learning) from the book "Dive into Deep Learning "



61

IPYTHON

- Interactive Python started in 2001 as an enhanced Python interpreter
- Developed by Fernando Perez as "Tools for the entire life cycle of research computing"
- If Python is Engine, IPython as the interactive control panel.
- Closely tied with the Jupyter project which provides browser based notebook
- Two modes
  - IPython shell (Anaconda prompt -> Ipython)
  - Jupyter notebook (Anaconda prompt -> jupyter ntoebook)

62

IPYTHON FEATURES

- Refer to online notebooks:
  - https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/01.00-IPython-Beyond-Normal-Python.ipynb

63

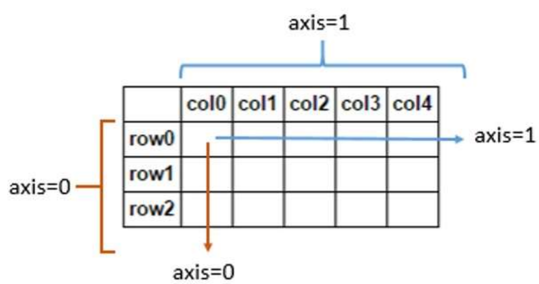
DATA AS NUMBERS

- Datasets can come from a wide range of sources and formats
  - E.g. documents, images, sound clips, numerical measurements
- Data is fundamentally array of numbers
- Digital images are 2D arrays of numbers representing pixel brightness across the area
- Sound clips are 1D arrays of intensity versus time
- Text can be converted in various ways into numerical representations
- First step in making data analyzable is to transform it into arrays of numbers
- Both, NumPy and Pandas package efficiently store and manipulate numerical arrays

64

NUMPY

- Stands for Numerical Python
- Efficient interface to store and operate on dense data buffers
- NumPy arrays are similar to Python's built-in list type, but provide much more efficient storage and data operations for larger arrays
- Form the core of data science tools in Python



65

NUMPY FEATURES

- Refer to online notebooks:
  - https://github.com/jakevdp/PythonDataScienceHandbook/blob/8a34a4f653bdbdc01415a94dc20d4e9b97438965/notebooks/02.00-Introduction-to-NumPy.ipynb

66

TENSORFLOW

- Deep learning framework released by Google in November 2015
- Deep learning does a wonderful job in pattern recognition, especially in the context of images, sound, speech, language, and time-series data.
- Installation: https://www.tensorflow.org/install/
- Examples and tutorials: https://github.com/tensorflow/examples

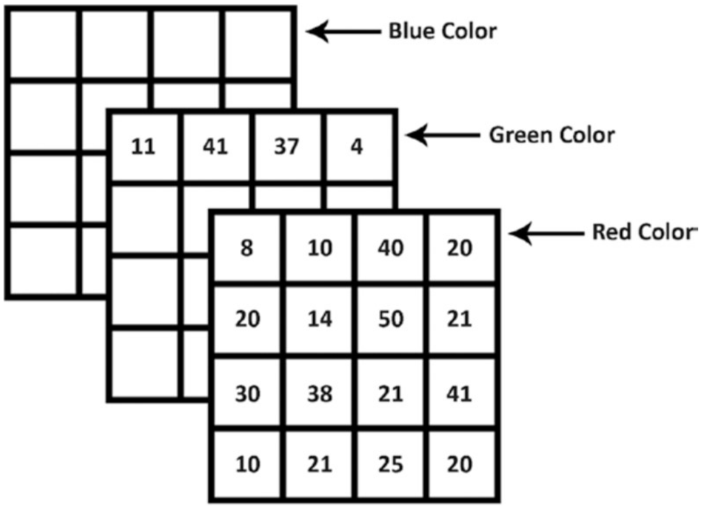
67

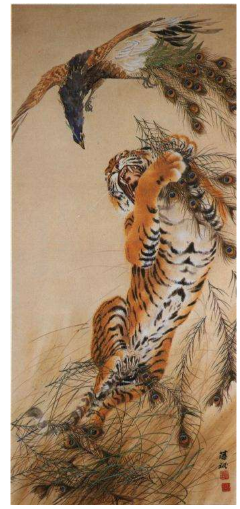
WHAT IS A TENSOR

- A tensor is a mathematical object and a generalization of scalars, vectors and matrices.
- A tensor can be represented as a multidimensional array.
- A tensor with zero rank (order) is a scalar.
- A tensor with rank 1 is a vector/array.
- Matrix is a tensor of rank 2.
  - 5: This is a rank 0 tensor; this is a scalar with shape [ ].
  - [2.,5., 3.]: This is a rank 1 tensor; this is a vector with shape [3].
  - [[1., 2., 7.], [3., 5., 4.]]: This is a rank 2 tensor; it is a matrix with shape [2, 3].
  - [[[1., 2., 3.]], [[7., 8., 9.]]]: This is a rank 3 tensor with shape [2, 1, 3].

68

CONVERT IMAGE TO A TENSOR





69

SCALARS

- Scalar is the value consisting of just one numerical quantity
- A scalar is represented by a tensor with just one element.
- In the next snippet, we instantiate two scalars and perform some familiar arithmetic operations with them, namely addition, multiplication, division, and exponentiation

```
import tensorflow as tf

x = tf.constant([3.0])
y = tf.constant([2.0])

x + y, x * y, x / y, x**y
```

```
(<tf.Tensor: shape=(1,), dtype=float32, numpy=array([5.], dtype=float32)>,
 <tf.Tensor: shape=(1,), dtype=float32, numpy=array([6.], dtype=float32)>,
 <tf.Tensor: shape=(1,), dtype=float32, numpy=array([1.5], dtype=float32)>,
 <tf.Tensor: shape=(1,), dtype=float32, numpy=array([9.], dtype=float32)>)
```

70

VECTORS

- A list of scalar values. We call these values the *elements* (*entries* or *components*) of the vector
- In ML vectors represent examples from the dataset and their values hold some real-world significance
- E.g., if we were training a model to predict the risk that a loan defaults, we might associate each applicant with a vector whose components correspond to their income, length of employment, number of previous defaults, and other factors.

```
x = tf.range(4)
x
```

```
<tf.Tensor: shape=(4,), dtype=int32, numpy=array([0, 1, 2, 3], dtype=int32)>
```

- In math, a vector x can be written as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

71

LENGTH, DIMENSIONALITY AND SHAPE

- The length of a tensor is given by Python's built-in len() function

```
len(x)
```

```
4
```

- The shape is a tuple that lists the length (dimensionality) along each axis of the tensor. For tensors with just one axis, the shape has just one element

```
x.shape
```

```
TensorShape([4])
```

72

MATRICES

- Matrices generalize vectors from order one to order two
- For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, the shape of $\mathbf{A}$ is (m , n) or $m \times n$. Specifically, when a matrix has the same number of rows and columns, its shape becomes a square; thus, it is called a square matrix.
- We can create an $m \times n$ matrix by specifying a shape with two components m and n when calling any of our favorite functions for instantiating a tensor.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

```
A = tf.reshape(tf.range(20), (5, 4))
A
```

```
<tf.Tensor: shape=(5, 4), dtype=int32, numpy=
array([[ 0,  1,  2,  3],
       [ 4,  5,  6,  7],
       [ 8,  9, 10, 11],
       [12, 13, 14, 15],
       [16, 17, 18, 19]], dtype=int32)>
```

73

MATRIX TRANSPOSE

- When we exchange a matrix's rows and columns, the result is called the *transpose* of the matrix
- We signify a matrix $\mathbf{A}$'s transpose by $\mathbf{A}^\top$ and if $\mathbf{B}=\mathbf{A}^\top$, then $b_{ij}=a_{ji}$ for any i and j
- A symmetric matrix $\mathbf{A}$ is equal to its transpose $\mathbf{A}^\top$

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}.$$

```
tf.transpose(A)
```

```
<tf.Tensor: shape=(4, 5), dtype=int32, numpy=
array([[ 0,  4,  8, 12, 16],
       [ 1,  5,  9, 13, 17],
       [ 2,  6, 10, 14, 18],
       [ 3,  7, 11, 15, 19]], dtype=int32)>
```

```
B = tf.constant([[1, 2, 3], [2, 0, 4], [3, 4, 5]])
B == tf.transpose(B)
```

```
<tf.Tensor: shape=(3, 3), dtype=bool, numpy=
array([[ True,  True,  True],
       [ True,  True,  True],
       [ True,  True,  True]])>
```

74

BASIC PROPERTIES OF TENSOR ARITHMETIC

- Given any two tensors with the same shape, the result of any binary elementwise operation will be a tensor of that same shape.

```
A = tf.reshape(tf.range(20, dtype=tf.float32), (5, 4))
B = A  # No cloning of `A` to `B` by allocating new memory
A, A + B
```

```
(<tf.Tensor: shape=(5, 4), dtype=float32, numpy=
 array([[ 0.,  1.,  2.,  3.],
        [ 4.,  5.,  6.,  7.],
        [ 8.,  9., 10., 11.],
        [12., 13., 14., 15.],
        [16., 17., 18., 19.]], dtype=float32)>,
 <tf.Tensor: shape=(5, 4), dtype=float32, numpy=
 array([[ 0.,  2.,  4.,  6.],
        [ 8., 10., 12., 14.],
        [16., 18., 20., 22.],
        [24., 26., 28., 30.],
        [32., 34., 36., 38.]], dtype=float32)>)
```

75

HADAMARD PRODUCT

- Elementwise multiplication of two matrices is called their Hadamard product (math notation $\odot$)

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix}.$$

```
A * B
```

```
<tf.Tensor: shape=(5, 4), dtype=float32, numpy=
array([[  0.,   1.,   4.,   9.],
       [ 16.,  25.,  36.,  49.],
       [ 64.,  81., 100., 121.],
       [144., 169., 196., 225.],
       [256., 289., 324., 361.]], dtype=float32)>
```

76

REDUCTION – SUM OF TENSOR ELEMENTS

```
x = tf.range(4, dtype=tf.float32)
x, tf.reduce_sum(x)
```

```
(<tf.Tensor: shape=(4,), dtype=float32, numpy=array([0., 1., 2., 3.], dtype=flo
at32)>,
 <tf.Tensor: shape=(), dtype=float32, numpy=6.0>)
```

We can express sums over the elements of tensors of arbitrary shape. For example, the sum of the elements of an $m \times n$ matrix $\mathbf{A}$ could be written $\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}$.

```
A.shape, tf.reduce_sum(A)
```

```
(TensorShape([5, 4]), <tf.Tensor: shape=(), dtype=float32, numpy=190.0>)
```

77

REDUCTION – ROWWISE (AXIS=0), COLUMNWISE (AXIS=1)

```
A_sum_axis0 = tf.reduce_sum(A, axis=0)
A_sum_axis0, A_sum_axis0.shape
```

```
(<tf.Tensor: shape=(4,), dtype=float32, numpy=array([40., 45., 50., 55.], dtype
=float32)>,
 TensorShape([4]))
```

```
A_sum_axis1 = tf.reduce_sum(A, axis=1)
A_sum_axis1, A_sum_axis1.shape
```

```
(<tf.Tensor: shape=(5,), dtype=float32, numpy=array([ 6., 22., 38., 54., 70.],
dtype=float32)>,
 TensorShape([5]))
```

Reducing a matrix along both rows and columns via summation is equivalent to summing up all the elements of the matrix.

```
tf.reduce_sum(A, axis=[0, 1])  # Same as `tf.reduce_sum(A)`
```

```
<tf.Tensor: shape=(), dtype=float32, numpy=190.0>
```

78

MEAN AKA AVERAGE

```
tf.reduce_mean(A), tf.reduce_sum(A) / tf.size(A).numpy()
```

```
(<tf.Tensor: shape=(), dtype=float32, numpy=9.5>,
 <tf.Tensor: shape=(), dtype=float32, numpy=9.5>)
```

Likewise, the function for calculating the mean can also reduce a tensor along the specified axes.

```
tf.reduce_mean(A, axis=0), tf.reduce_sum(A, axis=0) / A.shape[0]
```

```
(<tf.Tensor: shape=(4,), dtype=float32, numpy=array([ 8.,  9., 10., 11.], dtype
=float32)>,
 <tf.Tensor: shape=(4,), dtype=float32, numpy=array([ 8.,  9., 10., 11.], dtype
=float32)>)
```

79

NON-REDUCTION SUM, CUMULATIVE SUM

- Keep the number of axes unchanged when invoking the function for calculating the sum or mean

```
sum_A = tf.reduce_sum(A, axis=1, keepdims=True)
sum_A
```

```
<tf.Tensor: shape=(5, 1), dtype=float32, numpy=
array([[ 6.],
       [22.],
       [38.],
       [54.],
       [70.]], dtype=float32)>
```

- Calculate cumulative sum of elements of A along some axis (e.g. axis=0 by row)

```
tf.cumsum(A, axis=0)
```

```
<tf.Tensor: shape=(5, 4), dtype=float32, numpy=
array([[ 0.,  1.,  2.,  3.],
       [ 4.,  6.,  8., 10.],
       [12., 15., 18., 21.],
       [24., 28., 32., 36.],
       [40., 45., 50., 55.]], dtype=float32)>
```

80

DOT PRODUCT

Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, their *dot product* $\mathbf{x}^\top \mathbf{y}$ (or $\langle \mathbf{x}, \mathbf{y} \rangle$) is a sum over the products of the elements at the same position: $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$.

```
y = tf.ones(4, dtype=tf.float32)
x, y, tf.tensordot(x, y, axes=1)
```

```
(<tf.Tensor: shape=(4,), dtype=float32, numpy=array([0., 1., 2., 3.], dtype=flo
at32)>,
 <tf.Tensor: shape=(4,), dtype=float32, numpy=array([1., 1., 1., 1.], dtype=flo
at32)>,
 <tf.Tensor: shape=(), dtype=float32, numpy=6.0>)
```

Note that we can express the dot product of two vectors equivalently by performing an elementwise multiplication and then a sum:

```
tf.reduce_sum(x * y)
```

```
<tf.Tensor: shape=(), dtype=float32, numpy=6.0>
```

81

MATRIX-VECTOR PRODUCTS

- Let A be matrix represented using its row vectors
- The matrix-vector product $\mathbf{Ax}$ is simply a column vector of length m, whose i^{th} element is the dot product $\mathbf{a}_i^\top \mathbf{x}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix}, \qquad \mathbf{Ax} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix}.$$

```
A.shape, x.shape, tf.linalg.matvec(A, x)
```

```
(TensorShape([5, 4]),
 TensorShape([4]),
 <tf.Tensor: shape=(5,), dtype=float32, numpy=array([ 14.,  38.,  62.,  86., 110.], dtype=float32)>)
```

MATRIX-MATRIX MULTIPLICATION

```
B = tf.ones((4, 3), tf.float32)
tf.matmul(A, B)
```

```
<tf.Tensor: shape=(5, 3), dtype=float32, numpy=
array([[ 6.,  6.,  6.],
       [22., 22., 22.],
       [38., 38., 38.],
       [54., 54., 54.],
       [70., 70., 70.]], dtype=float32)>
```

82

NORMS

- Informally, the norm of a vector tells us how *big* a vector is in the magnitude
- The norm must be non-negative

Euclidean distance is a norm: specifically it is the L_2 norm. Suppose that the elements in the n-dimensional vector $\mathbf{x}$ are $x_1, \ldots, x_n$. The L_2 *norm* of $\mathbf{x}$ is the square root of the sum of the squares of the vector elements:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2},$$

where the subscript 2 is often omitted in L_2 norms, i.e., $\|\mathbf{x}\|$ is equivalent to $\|\mathbf{x}\|_2$. In code, we can calculate the L_2 norm of a vector as follows.

```
u = tf.constant([3.0, -4.0])
tf.norm(u)
```

```
<tf.Tensor: shape=(), dtype=float32, numpy=5.0>
```

83

L1 NORM

The L_1 *norm* is expressed as the sum of the absolute values of the vector elements:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| .$$

As compared with the L_2 norm, it is less influenced by outliers. To calculate the L_1 norm, we compose the absolute value function with a sum over the elements.

```
tf.reduce_sum(tf.abs(u))
```

```
<tf.Tensor: shape=(), dtype=float32, numpy=7.0>
```

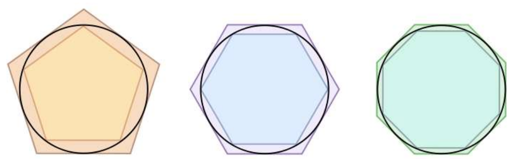
Both the L_2 norm and the L_1 norm are special cases of the more general L_p *norm*:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} .$$

84

CALCULUS – METHOD OF EXHAUSTION

- The method of exhaustion is a method of finding the area of a shape by inscribing inside it a sequence of polygons whose areas converge to the area of the containing shape
- If the sequence is correctly constructed, the difference in area between the n^{th} polygon and the containing shape will become arbitrarily small as n becomes large
- As this difference becomes arbitrarily small, the possible values for the area of the shape are systematically "exhausted" by the lower bound areas successively established by the sequence members
- Euclid used this method to prove certain propositions
  - The area of circles is proportional to the square of their diameters
  - The volumes of two tetrahedra of the same height are proportional to the areas of their triangular bases
- Integral calculus originated from the method of exhaustion



85

OPTIMIZATION AND GENERALIZATION IN MACHINE LEARNING MODELS

- In machine learning, we train models, updating them successively so that they get better and better as they see more and more data
- Getting better means minimizing a loss function, a score that answers the question "how bad is our model?"
- We really care about is producing a model that performs well on data that we have never seen before
- But we can only fit the model to data that we can actually see
- The task of fitting models can be decomposed into two key concerns:
  - i) Optimization: the process of fitting our models to observed data;
  - ii) Generalization: the mathematical principles and practitioners' wisdom that guide as to how to produce models whose validity extends beyond the exact set of data examples used to train them

86

DERIVATIVE AND DIFFERENTIATION

- In machine learning, loss functions are differentiable with respect to model parameters
- For each parameter, determine how rapidly the loss would increase or decrease if we make a small change to the parameter
- For a function *f* with input and output as scalars, the derivative of *f* is defined as:

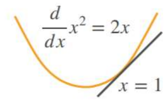
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

y	a	x^n	$\exp(x)$	$\log(x)$	$\sin(x)$
$\frac{dy}{dx}$	0	nx^{n-1}	$\exp(x)$	$\frac{1}{x}$	$\cos(x)$

a is not a function of x

y	$u+v$	uv	$y=f(u),\ u=g(x)$
$\frac{dy}{dx}$	$\frac{du}{dx}+\frac{dv}{dx}$	$\frac{du}{dx}v+\frac{dv}{dx}u$	$\frac{dy}{du}\frac{du}{dx}$

Derivative is the slope of the tangent line



The slope of the tangent line is 2

87

EXAMPLE: $f(x) = 3x^2 - 4x$

```
%matplotlib inline
from d2l import tensorflow as d2l
from IPython import display
import numpy as np

def f(x):
    return 3 * x ** 2 - 4 * x
```

```
def numerical_lim(f, x, h):
    return (f(x + h) - f(x)) / h

h = 0.1
for i in range(5):
    print(f'h={h:.5f}, numerical limit={numerical_lim(f, 1, h):.5f}')
    h *= 0.1
```

```
h=0.10000, numerical limit=2.30000
h=0.01000, numerical limit=2.03000
h=0.00100, numerical limit=2.00300
h=0.00010, numerical limit=2.00030
h=0.00001, numerical limit=2.00003
```

88

DIFFERENTIATION RULES

Suppose that functions f and g are both differentiable and C is a constant, we have:

the *constant multiple rule*

$$\frac{d}{dx}[Cf(x)] = C\frac{d}{dx}f(x),$$

the *sum rule*

$$\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}f(x) + \frac{d}{dx}g(x),$$

the *product rule*

$$\frac{d}{dx}[f(x)g(x)] = f(x)\frac{d}{dx}[g(x)] + g(x)\frac{d}{dx}[f(x)],$$

and the *quotient rule*

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)\frac{d}{dx}[f(x)] - f(x)\frac{d}{dx}[g(x)]}{[g(x)]^2}.$$

89

PARTIAL DERIVATIVE

In machine learning, functions often depend on *many* variables. Thus, we need to extend the ideas of differentiation to these *multivariate* functions.

Let $y = f(x_1, x_2, \ldots, x_n)$ be a function with n variables. The *partial derivative* of y with respect to its i^{th} parameter x_i is

$$\frac{\partial y}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}.$$

To calculate $\frac{\partial y}{\partial x_i}$, we can simply treat $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ as constants and calculate the derivative of y with respect to x_i. For notation of partial derivatives, the following are equivalent:

$$\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial x_i} = f_{x_i} = f_i = D_i f = D_{x_i} f.$$

Let $f(x, y) = y^3x^2$. Calculate $\frac{\partial f}{\partial x}(x, y)$.

$$\frac{\partial f}{\partial x}(x, y) = 2y^3x.$$

90

GRADIENTS

- A gradient is a vector whose components are the partial derivatives of a multivariate function with respect to all its variables
- Example: Say that we have a function, f(x,y) = 3x²y. Our partial derivatives are:

$$\frac{\partial f(x,y)}{\partial x} = \frac{\partial}{\partial x}3x^2y = 6yx \qquad \frac{\partial f(x,y)}{\partial y} = \frac{\partial}{\partial y}3x^2y = 3x^2$$

- If we organize these partials into a horizontal vector, we get the gradient of f(x,y), or ∇ f(x,y):

$$[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}] = [6yx, 3x^2]$$

- Gradient vector points to the direction of greatest increase of a function
- Gradient is zero at local maximum or local minimum (as there is no single direction of increase)

Explanatory video: https://www.youtube.com/watch?v=GkB4vW16QHI
Further info: https://betterexplained.com/articles/vector-calculus-understanding-the-gradient/

91

CHAIN RULE

- The chain rule enables us to differentiate composite functions
- Suppose that functions $y=f(u)$ and $u=g(x)$ are both differentiable, then the chain rule states that:

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

Suppose that the differentiable function y has variables $u_1, u_2, \ldots, u_m$, where each differentiable function u_i has variables $x_1, x_2, \ldots, x_n$. Note that y is a function of $x_1, x_2, \ldots, x_n$. Then the chain rule gives

$$\frac{dy}{dx_i} = \frac{dy}{du_1}\frac{du_1}{dx_i} + \frac{dy}{du_2}\frac{du_2}{dx_i} + \cdots + \frac{dy}{du_m}\frac{du_m}{dx_i}$$

for any $i = 1, 2, \ldots, n$.

92

CHAIN RULE EXAMPLE

Let $f(x) = 6x + 3$ and $g(x) = -2x + 5$. Use the chain rule to calculate $h'(x)$, where $h(x) = f(g(x))$.

Solution: The derivatives of f and g are

$$\begin{aligned} f'(x) &= 6 \\ g'(x) &= -2. \end{aligned}$$

According to the chain rule,

$$\begin{aligned} h'(x) &= f'(g(x))g'(x) \\ &= f'(-2x+5)(-2) \\ &= 6(-2) = -12. \end{aligned}$$

Let $f(x) = e^x$ and $g(x) = 4x$. Use the chain rule to calculate $h'(x)$, where $h(x) = f(g(x))$.

Solution: The derivative of the exponential function with base e is just the function itself, so $f'(x) = e^x$. The derivative of g is $g'(x) = 4$. According to the chain rule,

$$\begin{aligned} h'(x) &= f'(g(x))g'(x) \\ &= f'(4x) \cdot 4 \\ &= 4e^{4x}. \end{aligned}$$

- For more examples see: https://mathinsight.org/chain_rule_simple_examples

93