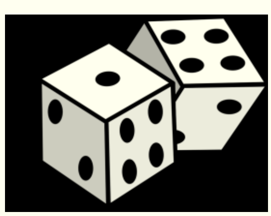
Statistical Inference

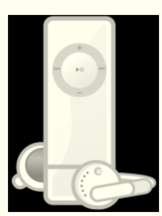
Probability

Behnam Bahrak

Random Process

- In a random process we know what outcomes could happen, but we don't know which particular outcome will happen.
- It can be helpful to model a process as random even if it is not truly random.





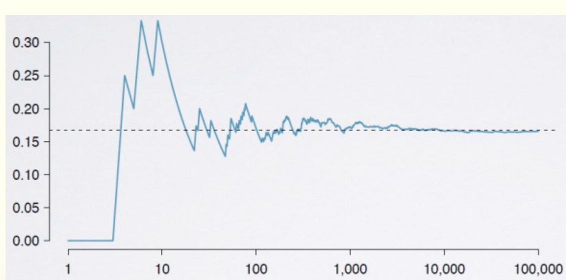
Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$
- Frequentist interpretation:
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- Bayesian interpretation:
  - A Bayesian interprets probability as a subjective degree of belief: For the same event, two separate people could have different viewpoints and so assign different probabilities.
  - Largely popularized by revolutionary advance in computational technology and methods during the last twenty years.

Law of Large Numbers

- Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome ($\hat{p}_n$) converges to the probability of that outcome (p).

$\hat{p}_n$



n (number of rolls)

Gambler's Fallacy

- When tossing a fair coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss?
  - 0.5, less than 0.5, or more than 0.5? H H H H H H H H H H ?
- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.
  - P(H on 11^{th} toss) = P(T on 11^{th} toss) = 0.5
- The coin is not due for a tail.
- The common misunderstanding of the LLN is that random processes are supposed to compensate for whatever happened in the past; this is just not true and is also called gambler's fallacy (or law of averages).

Gambler's Fallacy

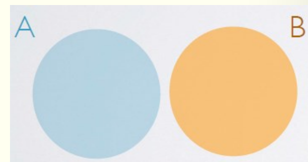
- On August 18, 1913, at a casino in Monte Carlo, black came up a record twenty-six times in succession [in roulette] … There was a near panicky rush to bet on red, beginning about the time black had come up a phenomenal fifteen times.

  -- Huff and Geis, How to Take a Chance

- Probability of 26 consecutive reds
  - 1 in 66.6 million
- Probability of 26 consecutive reds when previous 25 rolls were red
  - Almost 1/2

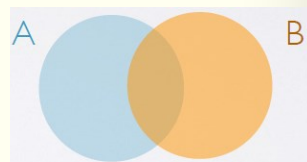
Disjoint (Mutually Exclusive) Events

- Disjoint (mutually exclusive) events cannot happen at the same time.
  - the outcome of a single coin toss cannot be a head and a tail.
  - a student can't both fail and pass a class.
  - a single card drawn from a deck cannot be an ace and a queen.



$P(A \cap B) = 0$

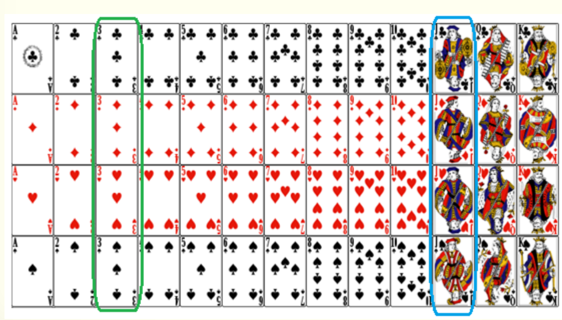
- Non-disjoint events can happen at the same time.
  - A student can get an A in Stats and A in Graph in the same semester



$P(A \cap B) \neq 0$

Union of Disjoint Events

- What is the probability of drawing a Jack or a three from a well shuffled full deck of cards?



$P(J \text{ or } 3) =$

$P(J) + P(3) =$

$4/52 + 4/52 =$

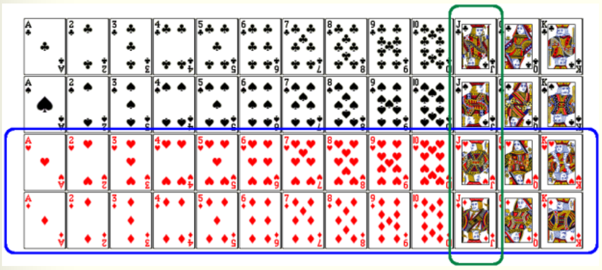
≈ 0.154

- For disjoint events A and B:

$P(A \cup B) = P(A) + P(B)$

Union of Non-disjoint Events

- What is the probability of drawing a jack or a red card from a well shuffled full deck?



$P(J \text{ or } red) =$

$P(red) + P(j) - P(j \, and \, red)$

$= 26/52 + 4/52 - 2/52$

≈ 0.538

- For non-disjoint events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Sample Space

- A sample space is a collection of all possible outcomes of a trial.
- A couple has one kid, what is the sample space for the gender of this kid?

S = {M, F}

- A couple has two kids, what is the sample space for the gender of these kids?

S = {MM, MF, FM, FF}

- Events are subsets of the sample space.

Probability Distributions

- A probability distribution lists all possible outcomes in the sample space, and the probabilities with which they occur.

one toss	H	T
probability	0.5	0.5

two tosses	HH	HT	TH	TT
probability	0.25	0.25	0.25	0.25

- *Kolmogorov rules*
  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1

Complementary Events

- Complementary events are two mutually exclusive events whose probabilities add up to 1.

complementary

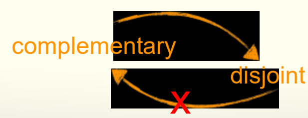
one toss	H	T
probability	0.5	0.5

complementary

two tosses	HH	[redacted]
probability	0.25	[redacted]

Disjoint vs. Complementary

- Do the sum of probabilities of two disjoint outcomes always add up to 1?
  - Not necessarily, there may be more than 2 outcomes in the sample space.
- Do the sum of probabilities of two complementary outcomes always add up to 1?
  - Yes, that's the definition of complementary



Independence

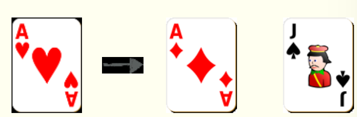
- Two random processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

1st toss 2nd toss

$P(H) = 0.5$ $P(T) = 0.5$

outcomes of two tosses of a coin are independent

1st draw 2nd draw



$P(A) = \frac{3}{51}$ $P(J) = \frac{4}{51}$

outcomes of two draws from a deck of cards (without replacement) are dependent

Question

Between January 9-12, 2013, SurveyUSA interviewed a random sample of 500 NC residents asking them whether they think widespread gun ownership protects law abiding citizens from crime, or makes society more dangerous. 58% of all respondents said it protects citizens. 67% of White respondents, 28% of Black respondents, and 64% of Hispanic respondents shared this view. Which of the below is true?

Opinion on gun ownership and race ethnicity are most likely ---------.

(a) Complimentary
(b) Mutually exclusive
(c) Independent
(d) Dependent ✓
(e) Disjoint

Checking for Independence

- If P(A occurs, given that B is true) = P(A|B) = P(A), then A and B are independent.

P(protects citizens) = 0.58

P(protects citizens | White) = 0.67

P(protects citizens | Black) = 0.28

P(protects citizens | Hispanic) = 0.64

- P(protects citizens) varies by race/ethnicity, therefore opinion on gun ownership and race ethnicity are most likely dependent.

Determining dependence based on sample data

observed difference between conditional probabilities → dependence → hypothesis test

if difference is large, there is stronger evidence that the difference is real

if sample size is large, even a small difference can provide strong evidence of a real difference

Example

- We saw that P(protects citizens|White) = 0.67 and P(protects citizens|Hispanic) = 0.64. Under which condition would you be more convinced of a real difference between the proportions of Whites and Hispanics who think gun widespread gun ownership protects citizens? n = 500 or n = 50,000?

n = 50,000

Independent Events

- Product rule for independent events:
  If A and B are independent, $P(A \cap B) = P(A) \times P(B)$

Example: You toss a coin twice, what is the probability of getting two tails in a row?

P(two tails in a row) =
P(T on the 1st toss) × P(T on the 2nd toss) =
(1/2) × (1/2) = 1/4

- Note: If $A_1, A_2, \dots, A_k$ are independent:

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

Independence of more than 2 events

- n events $A_1, A_2, \dots, A_n$ are independent if and only if:

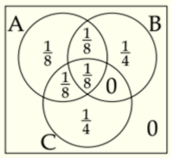
$$P\left(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_r}\right) = P\left(A_{k_1}\right)P\left(A_{k_2}\right) \dots P\left(A_{k_r}\right)$$

for all $\{k_1, k_2, \dots, k_r\} \subseteq \{1, 2, \dots, n\}$ with $r \le n$.

Example: Three events A, B, C are independent if we have:

1) $P(A \cap B) = P(A)P(B)$
2) $P(A \cap C) = P(A)P(C)$
3) $P(B \cap C) = P(B)P(C)$
4) $P(A \cap B \cap C) = P(A)P(B)P(C)$

Independence of Three Events



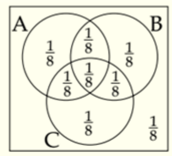
☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(B)$

☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(C)$

☒ $\frac{1}{8} \neq \frac{1}{2} \times \frac{1}{2} = P(B)P(C)$

☑ $\frac{1}{8} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = P(A)P(B)P(C)$

A and B are independent.
A and C are independent.



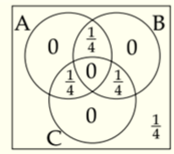
☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(B)$

☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(C)$

☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(B)P(C)$

☑ $\frac{1}{8} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = P(A)P(B)P(C)$

A, B and C are independent.



☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(B)$

☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(C)$

☑ $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(B)P(C)$

☒ $0 \neq \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = P(A)P(B)P(C)$

A and B are independent.
A and C are independent.
B and C are independent.

Disjoint vs. Independent

- Two events that are disjoint (mutually exclusive) cannot happen at the same time:

$$P(A \cap B) = 0$$

- Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other:

$$P(A|B) = P(A)$$

Infinite Monkey Theorem

- **Infinite monkey theorem**: a monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type any given text, such as the complete works of William Shakespeare.

Proof:

- Assume that the length of the text $\boldsymbol{T}$ (e.g. Shakespeare's Hamlet) is equal to m bits. We divide the infinite sequence to non-overlapping m-bit subsequences:

m = 4: 1101000100001111110101111000101010111100000111100...

- The probability of one of these subsequences being equal to $\boldsymbol{T}$ is $1/2^m$, thus considering the independence of these subsequences the probability that none of these subsequences is equal to $\boldsymbol{T}$ is:

$$\lim_{n\to\infty}\left(1-\frac{1}{2^m}\right)^{n/m}=0$$

Independence Fallacy

- In November 1999, Sally Clark was found guilty of the murder of her two infant sons.
- The defence argued that the children had died of sudden infant death syndrome (SIDS).

- The prosecution case relied on flawed statistical evidence presented by paediatrician Professor Sir Roy Meadow, who testified that the chance of two children from an affluent family suffering SIDS was 1 in 73 million:

$$\frac{1}{8500}\times\frac{1}{8500}=\frac{1}{72250000}$$

- The Royal Statistical Society later issued a statement arguing that there was no statistical basis for Meadow's claim, and expressed concern at the "misuse of statistics in the courts".

Birthday Problem

- How many people in a room does it take for it to have at least a 50% chance that two of them share a birthday?
- It's easier to find the probability of the complement event.
- Assume that there are n people in the room:

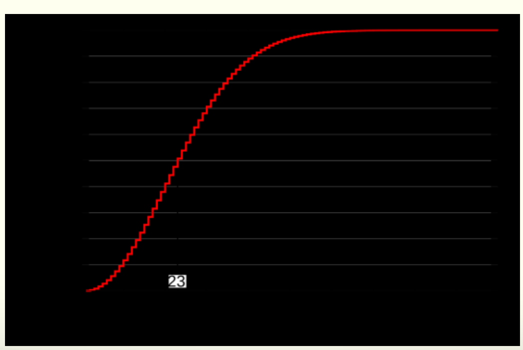
A = the event that no two people in the room have the same birthday

- The number of possible combinations of n people's birthdays is: 365^n
- The number of combinations where no two people have the same birthday:

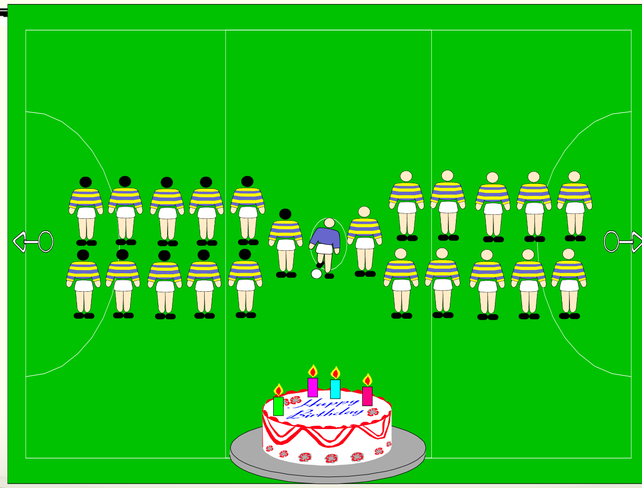
$$365 \times 364 \times \cdots \times (365 - n + 1) = \frac{365!}{(365-n)!}$$

Birthday Problem

- Therefore the probability that at least two people in the room have the same birthday: $1 - \dfrac{365!}{365^n(365-n)!}$



Birthday Paradox



Example Data Set

- ADOLESCENTS' UNDERSTANDING OF SOCIAL CLASS
  - Study examining teens' beliefs about social class
- **Sample:** 48 working class and 50 upper middle class 16-year-olds
- **Study Design:**
  - objective assignment to social class based on self-reported measures of both parents' occupation and education, and household income
  - subjective association based on survey questions

Contingency Table of the Data Set

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

Marginal Probability

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	[redacted]	[redacted]

- What is the probability that a student's objective social class position is upper middle class?

$$P(obj\ UMC) = 50/98 \approx 0.51$$

Joint Probability

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	[redacted]	45
	upper class	0	0	0
	Total	48	50	[redacted]

- What is the probability that a student's objective position and subjective identity are both upper middle class?

$$P(obj\ UMC \cap subj\ UMC) = 37/98 \approx 0.38$$

Conditional Probability

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	[redacted]	37	45
	upper class	0	0	0
	Total	[redacted]	50	98

- What is the probability that a student who is objectively in the working class associates with upper middle class?

$$P(subj\ UMC \mid obj\ WC) = 8/48 \approx 0.17$$

Conditional Probability

- We can also use the definition to find the conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- For the previous problem we have:

$$P(subj\ UMC \mid obj\ WC) = \frac{P(subj\ UMC \cap\ obj\ WC)}{P(obj\ WC)} = \frac{8/98}{48/98} = \frac{8}{48}$$

- General Product Rule:

$$P(A \cap B) = P(A|B)P(\mathrm{B})$$

Chain Rule

- Sometimes we write $P(A, B)$ or $P(AB)$ instead of $P(A \cap B)$ for simplicity.

- We can generalize the product rule to three events:

$$P(ABC) = P(AB|C)P(C) = P(A|BC)P(B|C)P(C)$$

- We can generalize this rule to n different events $A_1, \dots, A_n$:

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \dots P(A_n|A_1 A_2 \dots A_{n-1})$$

- This generalized property of conditional property is called the chain rule.

Independence and Conditional Probabilities

- Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.
  - Conceptually: Giving B doesn't tell us anything about A.
  - Mathematically: If events A and B are independent, $P(A \cap B) = P(A) \times P(B)$, thus:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Example

		major		
		social science	non-social science	Total
gender	female	30	20	50
	male	30	20	50
	Total	60	40	100

$P(SS) = 60 / 100 = 0.6$

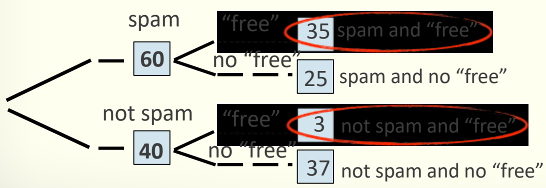
$P(SS \mid F) = 30 / 50 = 0.6$

$P(SS \mid M) = 30 / 50 = 0.6$

Probability Tree

➢ We often use a probability tree to compute conditional probabilities.

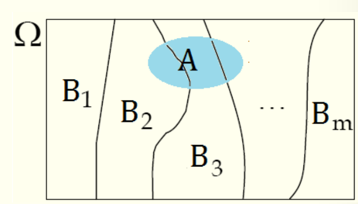
Example: You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word “free”. Of the rest, 3 contain the word “free”. If an email contains the word “free”, what is the probability that it is spam?



P(spam | “free”) $= \frac{35}{35+3} = 0.92$

Total Probability Theorem

➢ If $\{B_i : i = 1, 2, \ldots, m\}$ be a set of pairwise disjoint events, whose union is the entire sample space Ω, then:



$$P(A) = \sum_{i=1}^{m} P(A|B_i)P(B_i)$$

Proof:

$$P(A) = P(A \cap \Omega) = P(A \cap (\cup_{i=1}^{m} B_i)) = P(\cup_{i=1}^{m}(A \cap B_i))$$

$$= \sum_{i=1}^{m} P(A \cap B_i) = \sum_{i=1}^{m} P(A|B_i)P(B_i)$$

Bayes Theorem

- In many applications, we have $P(A|B)$, and we need to compute $P(B|A)$.
  - In such cases we use Bayes theorem.

- Bayes Theorem:

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

and in general:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{m} P(A|B_i)P(B_i)}$$

Thomas Bayes
(1701-1761)

Bayesian Interpretation of Probability

- A Bayesian interprets probability of an event as a subjective degree of belief about the event.
- Bayes theorem relates the conditional and marginal probabilities, i.e. it updates our knowledge or belief about event B after observing event A:

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

- $P(B)$: initial degree of belief in B (prior probability)
- $P(B|A)$: degree of belief after observing A (posterior probability)
- $P(A|B)/P(A)$: the support A provides for B

Monty Hall Problem

- Suppose you're on a game show, and you're given the choice of three doors:
  - Behind one door is a car, behind the others, goats.

- You pick a door, say No. 3, and the host, who knows what's behind the doors, opens another door, say No. 1, which has a goat.
- He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Monty Hall Problem

A_i: The car is behind door i

Prior probabilities: $P(A_1) = P(A_2) = P(A_3) = 1/3$

- Without loss of generality, assume that the contestant chooses door #3, and the host opens door #1:

  B_1: The event that the host opens door #1

$P(B_1 \mid A_1) = 0$, $P(B_1 \mid A_2) = 1$, $P(B_1 \mid A_3) = 1/2$

$$P(B_1) = P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2) + P(B_1|A_3)P(A_3)$$

$$= 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{1}{2}$$

Statistical Inference Behnam Bahrak

Monty Hall Problem

➢ According to the Bayes theorem, the posterior probabilities are:

$$P(A_1|B_1) = 0$$

$$P(A_2|B_1) = \frac{P(B_1|A_2)P(A_2)}{P(B_1)} = \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

$$P(A_3|B_1) = \frac{P(B_1|A_3)P(A_3)}{P(B_1)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

➢ It is to the contestant's advantage to switch his/her choice!

Example

➢ American Cancer Society estimates that about 2% of men above 40 have prostate cancer. The accuracy of a prostate cancer screening test is 95%, which means that the test result is positive for a person who has prostate cancer with a probability of 0.95, and it is negative for a healthy person also with probability 0.95. A patient's test result is positive, so his doctor says that he has prostate cancer with a probability of 95%. Is the doctor right?

H: The patient is healthy S: The patient has cancer
T^+: The test result is positive T^-: The test result is negative

$$P(S) = 0.02 \quad \Rightarrow P(H) = 1 - 0.02 = 0.98$$

Example

$P(T^+|S) = 0.95 \;\Rightarrow\; P(T^-|S) = 1 - 0.95 = 0.05$

$P(T^-|H) = 0.95 \;\Rightarrow\; P(T^+|H) = 1 - 0.95 = 0.05$

- We need to find $P(S|T^+)$:

$$P(S|T^+) = \frac{P(T^+|S)P(S)}{P(T^+|S)P(S) + P(T^+|H)P(H)}$$
$$= \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.05 \times 0.98} = 0.278$$

- The probability of the patient having cancer is only 27.8% and the doctor has confused $P(S|T^+)$ with $P(T^+|S)$.

Prosecutor's Fallacy

- The common mistake of conflating $P(A|B)$ with $P(B|A)$ is called the prosecutor's fallacy.

- This fallacy of statistical reasoning, typically used by the prosecution to argue for the guilt of a defendant during a criminal trial.

- Example: 90% of burglars have white cars. The defendant has a white car, so he is a burglar with a probability of 0.9
  - Conflating P(white car | burglar) with P(burglar | white car).

Sampling with Replacement

- When sampling *with replacement*, you put back what you just drew.
  - Imagine you have a bag with 5 red, 3 blue and 2 yellow chips in it. What is the probability that the first chip you draw is blue?

$$5 \text{ red}, 3 \text{ blue}, 2 \text{ yellow}$$

$$P(1^{st} \text{ chip } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

  - Suppose you did indeed pull a blue chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

$$1^{st} \text{ draw: } 5 \text{ red}, 3 \text{ blue}, 2 \text{ yellow}$$

$$2^{nd} \text{ draw: } 5 \text{ red}, 3 \text{ blue}, 2 \text{ yellow}$$

$$P(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } B) = \frac{3}{10} = 0.3$$

Sampling with Replacement

- Suppose you actually pulled a yellow chip in the first draw. If drawing with replacement, what is the probability of drawing a blue chip in the second draw?

$$1^{st} \text{ draw: } 5 \text{ red}, 3 \text{ blue}, 2 \text{ yellow}$$

$$2^{nd} \text{ draw: } 5 \text{ red}, 3 \text{ blue}, 2 \text{ yellow}$$

$$P(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } Y) = \frac{3}{10} = 0.3$$

- When drawing with replacement, probability of the second chip being blue does not depend on the color of the first chip since whatever we draw in the first draw gets put back in the bag: $P(B|Y) = P(B|B) = P(B)$
- Note: When drawing with replacement, draws are independent.

Sampling without Replacement

- In drawing without replacement you do not put back what you just drew:
  - Suppose you pulled a blue chip in the first draw. If drawing without replacement, what is the probability of drawing a blue chip in the second draw?

1^{st} draw: 5 ● , 3 ● , 2 ●
2^{nd} draw: 5 ● , 2 ● , 2 ●

$$Prob(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$$

- When drawing without replacement, the probability of the second chip being blue given the first was blue is not equal to the probability of drawing a blue chip in the first draw since the composition of the bag changes with the outcome of the first draw: $P(B|B) \neq P(B)$
- When drawing without replacement, draws are not independent.

Sampling from a Small Population

- Taking into account that if sampling is done with or without replacement is especially important when the sample sizes are small.
- If we were dealing with, say, 10,000 chips in a (giant) bag, taking out one chip of any color would not have as big an impact on the probabilities in the second draw.
- Sampling is usually done without replacement, thus when we are taking samples from a small population, since the draws are not independent, many of the methods that we cover in this course are not applicable.

Random Variable

- A random variable is a numeric quantity whose value depends on the outcome of a random event
  - We use a capital letter, like X, to denote a random variable
  - The values of a random variable are denoted with a lowercase letter, in this case x, e.g. $P(X = x)$.
- There are two types of random variables:
  - Discrete random variables often take only integer values
    - Example: Number of credit hours, face of dice
  - Continuous random variables take real (decimal) values
    - Example: Cost of books this term, time you arrive to the class

Example

- Assume that random variable X defines the number of heads in throwing 3 coins:

Event	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0

$$P(X=3) = \frac{1}{8}, \qquad P(X \leq 1) = \frac{4}{8}$$

Probability Mass Function (or PMF) of X:

X	0	1	2	3
$P(X)$	1/8	3/8	3/8	1/8

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the expected value (mean), and it is a weighted average of the possible outcomes:

$$\mu = E(X) = \sum_{i=1}^{k} x_i \, P\{X = x_i\}$$

- Example: X = the number of heads in throwing 3 coins

$$\mu = E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

Expectation Manipulation

- A school has 3 classes with 5, 10 and 150 students
- Randomly choose a <u>class</u> with equal probability
  - X = size of chosen class
- What is E[X]?

  E[X] = 5 (1/3) + 10 (1/3) + 150 (1/3) = 165/3 = 55
- Randomly choose a <u>student</u> with equal probability
  - Y = size of class that student is in
- What is E[Y]?

  E[Y] = 5 (5/165) + 10 (10/165) + 150 (150/165) = 22635/165 ≈ 137
- Note: E[Y] is students' perception of class size
  - But E[X] is what is usually reported by schools!

Betting on Different Beliefs

- Alice believes that Barcelona will win the Copa del Rey with probability 5/8. Bob believes Real Madrid will win the competition with probability 3/4.
- You bet Alice that you'll pay her $2000 if Barcelona wins and she'll pay you $3000 otherwise. Should Alice take the bet?

$$E[A] = 2000 \times \frac{5}{8} - 3000 \times \frac{3}{8} = 125$$

- You bet Bob that you'll pay him $2000 if Real Madrid wins and he'll pay you $3000 otherwise. Should Bob take the bet?

$$E[B] = 2000 \times \frac{3}{4} - 3000 \times \frac{1}{4} = 750$$

- As long as you find two people who have different beliefs about something, you can design such a scheme.

Variance

- We are also often interested in the variability (variance) in the values of a random variable:

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} \left(x_i - E(X)\right)^2 P(X = x_i)$$

- We define standard deviation of a random variable as:

$$\sigma = SD(X) = \sqrt{Var(X)}$$

- Example: X = the number of heads in throwing 3 coins

$$\sigma^2 = \left(0 - \frac{3}{2}\right)^2 \times \frac{1}{8} + \left(1 - \frac{3}{2}\right)^2 \times \frac{3}{8} + \left(2 - \frac{3}{2}\right)^2 \times \frac{3}{8} + \left(3 - \frac{3}{2}\right)^2 \times \frac{1}{8} = \frac{3}{4}$$

Cumulative Distribution Function

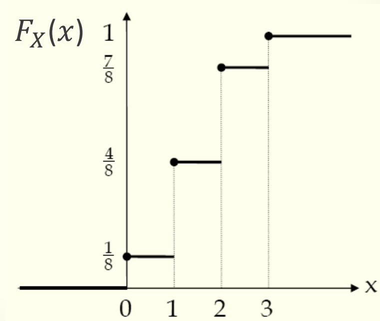
- Unlike a discrete random variable, the values that a continuous random variable X can take is uncountable, so we usually have:

$$P(X = x) = 0$$

- Thus we define another function which is called Cumulative Distribution Function (CDF) of a random variable X:

$$F_X(x) = P\{X \leq x\}$$

- For the 3 coins example we have:



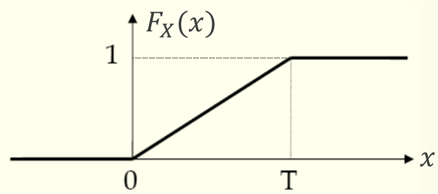
CDF Properties

1. $F_X(-\infty) = 0$
2. $F_X(+\infty) = 1$
3. $x_1 < x_2 :\ \ F_X(x_1) \leq F_X(x_2)$
4. $P\{X > x\} = 1 - F_X(x)$
5. $P\{x_1 < X \leq x_2\} = F_X(x_2) - F_X(x_1)$

Continuous Distribution

- We say random variable X is continuous if its cumulative distribution function is continuous everywhere:

X = random variable that represents the time of a phone call in the interval [0,T]

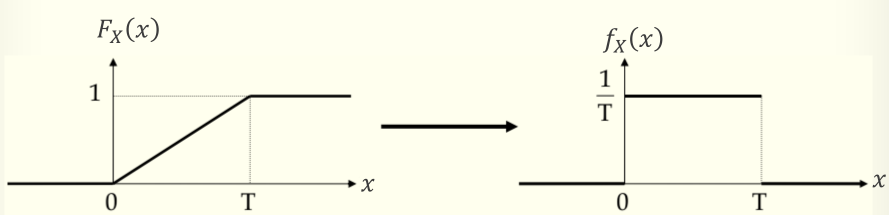


- We define probability density function (PDF) of a continuous random variable X as:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

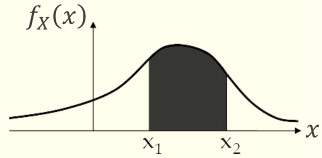
Probability Density Function

$$f(x) = \frac{dF(x)}{dx} \Rightarrow F(x) = \int_{-\infty}^{x} f(t)dt$$



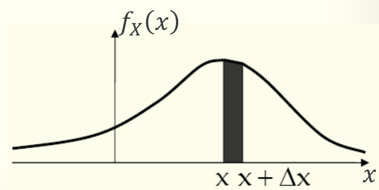
PDF Properties

1. $P\{x_1 \le X \le x_2\} = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X\,(x)\mathrm{d}x$



- The concept of density:

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P\{x \le X \le x + \Delta x\}}{\Delta x}$$



2. $f(x) \ge 0$
3. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$

Expected Value and Variance

- Expected value of a continuous random variable is defined as:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x\, f_X(x)dx$$

- Variance of a continuous random variable is defined as:

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2\, f_X(x)dx$$

- It is easy to show that: $Var(X) = E\left(X^2\right) - E(X)^2$

Linear Combinations

- A linear combination of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.

- Expectation is a linear operation, i.e. the expected value of a linear combination of random variables X and Y is given by:

$$E(aX + bY) = a\,E(X) + b\,E(Y)$$

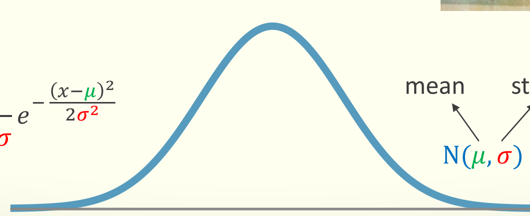
- The variance of a linear combination of two independent random variables is calculated as:

$$Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y)$$

Normal (Gaussian) Distribution

- The normal (or Gaussian) distribution is a very common continuous probability distribution which is very important in statistics.
  - Unimodal and symmetric
    - Bell curve

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

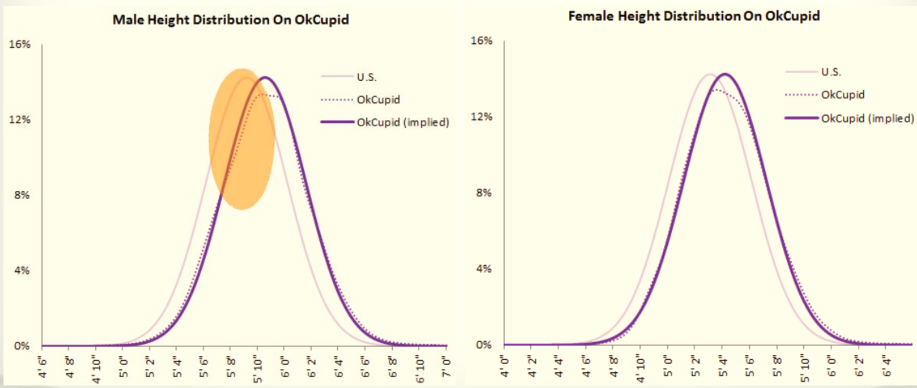


mean standard deviation

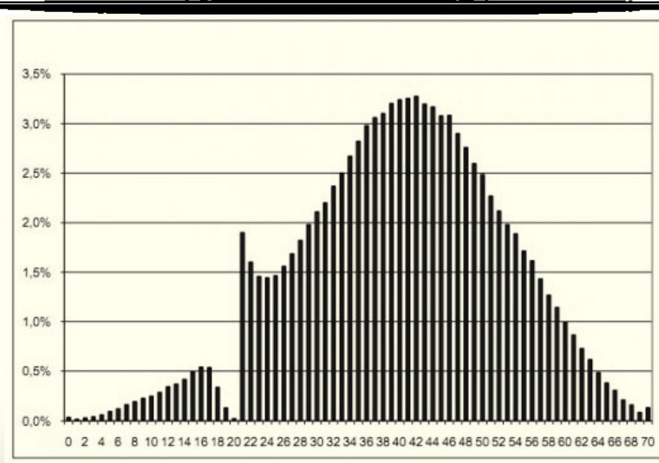
$N(\mu, \sigma)$

Normal Distribution in Real Life

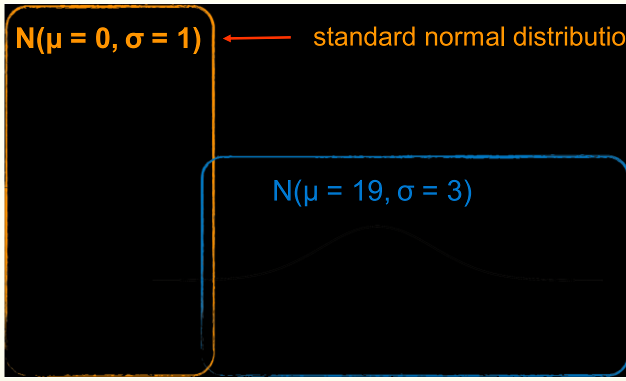
- Many random variables in nature has normal distribution.



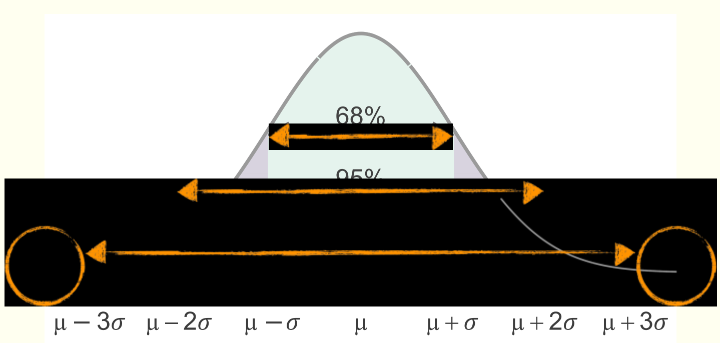
Final Grade Distribution in Poland



Normal Distribution



Empirical Rule (68 - 95 - 99.7% rule)



Question

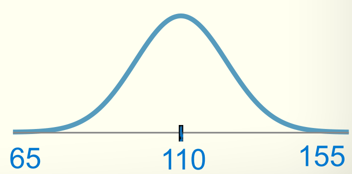
A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. He only reports 3 statistics, the mean = 110 beats per minute, the minimum = 65 beats per minute, and the maximum = 155 beats per minute. Which of the following is most likely to be the standard deviation of the distribution?

(a) 5 → $110 \pm (3\times5) = (95, 125)$

15 → $110 \pm (3\times15) = (65, 155)$

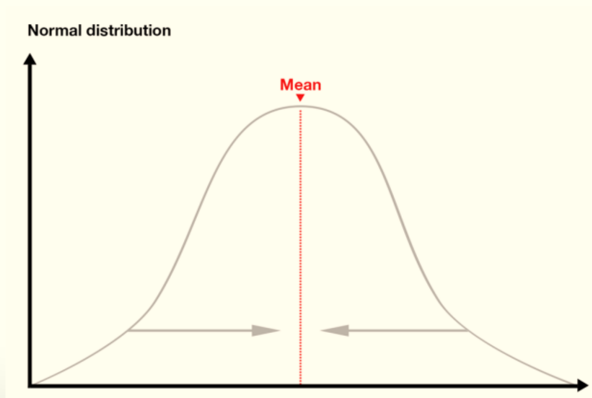
(c) 35 → $110 \pm (3\times35) = (5, 215)$

(d) 90 → $110 \pm (3\times90) = (-160, 380)$



Regression to the Mean

Why perfection rarely lasts?



Regression to the Mean

- Following an extreme random event, the next random event is likely to be less extreme.
- If you spin a fair roulette wheel 10 times and get 100% reds, that is an extreme event (probability = 1/1024).
- It is likely that in the next 10 spins, you will get fewer than 10 reds.
  - But the expected number is only 5
- So, if you look at the average of the 20 spins, it will be closer to the expected mean of 50% reds than to the 100% of the first 10 spins.
- Don't confuse "regression to the mean" with "gambler's fallacy"!

Standardizing with Z Scores

- Standardized (Z) score of an observation is the number of standard deviations it falls above or below the mean
- Z score of mean = 0
- Unusual observation: $|Z| > 2$
- Defined for distributions of any shape

$$Z = \frac{\text{observation} - \text{mean}}{\text{SD}}$$

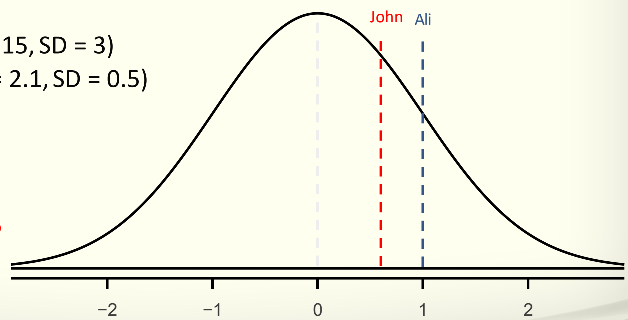
Example

- A university admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Ali, who has a GPA of 18 at UT, or John, who earned a GPA of 2.4 at MIT?

UT GPA ~ N(mean = 15, SD = 3)
MIT GPA ~ N(mean = 2.1, SD = 0.5)

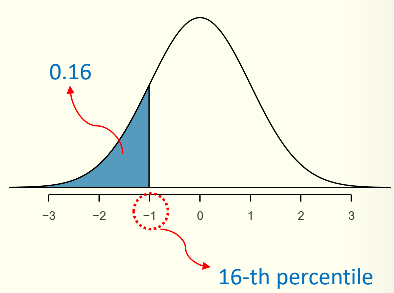
$$Z_{Ali} = \frac{18-15}{3} = 1$$

$$Z_{John} = \frac{2.4-2.1}{0.5} = 0.6$$



Percentile

- A percentile (or a centile) is a measure indicating the value below which a given percentage of observations in a group of observations fall.
- When the distribution is normal, Z scores can be used to calculate percentiles.



- Graphically, if the area below the PDF curve to the left of an observation, i.e. the CDF function, be equal to p, the observation is the $(100p)$-th percentile.

Computing Percentiles

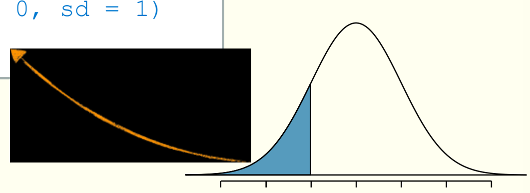
- Using Standard normal table:

Z	Second decimal place of Z: 0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Computing Percentiles

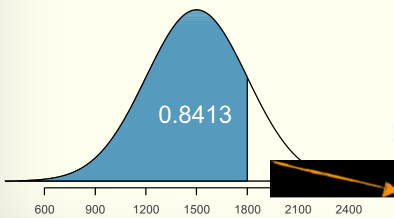
- Using R:

```
R
> pnorm(-1, mean = 0, sd = 1)
[1] 0.1586553
```



Example 1

R
```
> pnorm(18, mean = 15, sd = 3)
[1] 0.8413
```



$Z = \frac{18 - 15}{3} = 1$

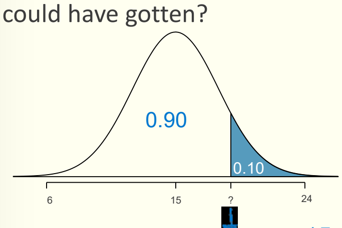
$P(Z < 1) = 0.8413$

1 - 0.8413 = 0.1587

Z		l decimal place of Z 0.01	0.02	0.
0.0		0.5040	0.5080	0.
0.1		0.5438	0.5478	0.
0.2		0.5832	0.5871	0.
0.8		0.7910	0.7939	0.
0.9		0.8186	0.8212	0.
1.1		0.8665	0.8686	0.

Example 2

A friend of yours tells you that her GPA is in the top 10% at UT. What is the lowest possible score she could have gotten?



$Z = 1.28 = \frac{X - 15}{3}$

$X = (1.28 \times 3) + 15 = 18.84$

Z	Second decimal place of Z 0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07		0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279		0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675		0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064		0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443		0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808		0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157		0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486		0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794		0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078		0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340		0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577		0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790		0.8830
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292		0.9319

R
```
> qnorm(0.90, 15, 3)
[1] 18.84465
```

Quantile

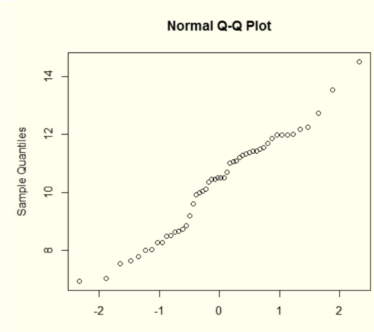
- **Quantiles** are cutpoints dividing a set of observations into equal sized groups, i.e. points in your data below which a certain proportion of your data fall.
  - 4-quantile: quartiles Q1 , Q2 , Q3
  - 100-quantile: percentile
    - The 0.95 quantile, or 95th percentile, of standard normal distribution is about 1.64
- Example: randomly generate data sample of size 200 from a standard normal distribution and find the quantiles for 0.01 to 0.99 using the `quantile` function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

- So we see that quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall.

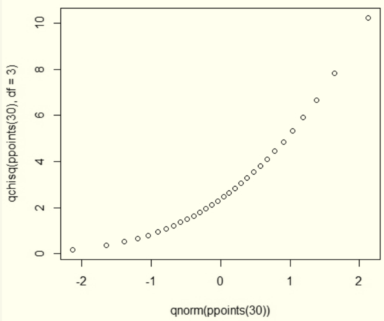
Q-Q Plots

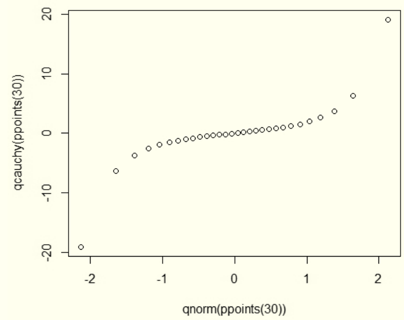
- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.
  - It's just a visual check, not an air-tight proof, so it is subjective.



- In R, there are two functions to create Q-Q plots:
  - `qqnorm(x)`: draws quantiles of data sample x vs. quantiles of standard normal
  - `qqplot(x,y)`: draws quantiles of data x vs. quantiles of data y

Q-Q Plots



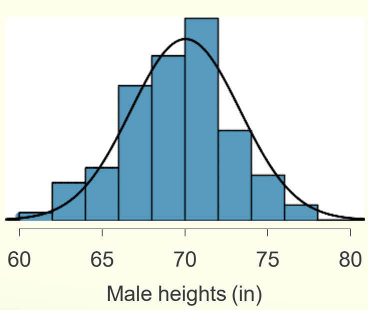


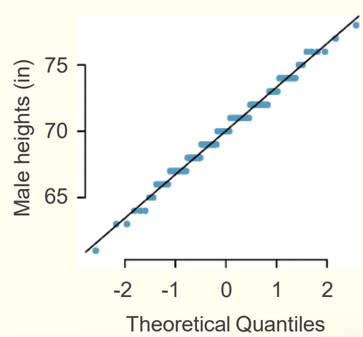
`qqplot(qnorm(ppoints(30)),qchisq(ppoints(30),df=3))`

`qqplot(qnorm(ppoints(30)), qcauchy(ppoints(30)))`

Normal Probability Plot

➢ A histogram and normal probability plot of a sample of 100 male heights.



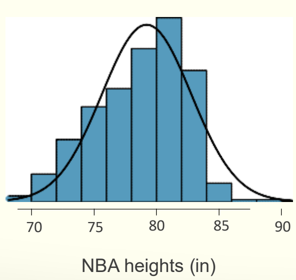


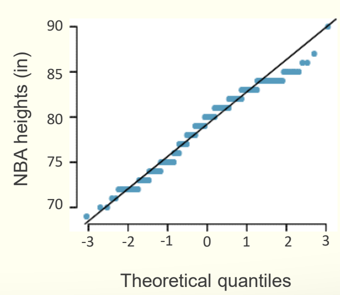
Anatomy of a Normal Probability Plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a standard normal distribution) on the x-axis.
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
- Constructing a normal probability plot requires tedious calculations, so we generally rely on software when making these plots.

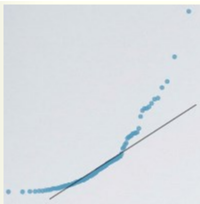
Example

- Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

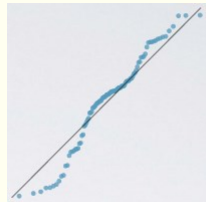




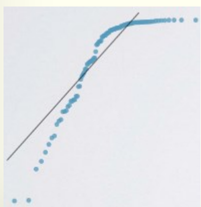
Normal Probability Plot and Skewness



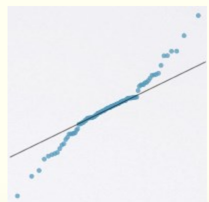
- Right skew
- Points bend up and to the left of the line.



- Short tails (narrower than the normal distribution)
- Points follow an S-shaped curve.



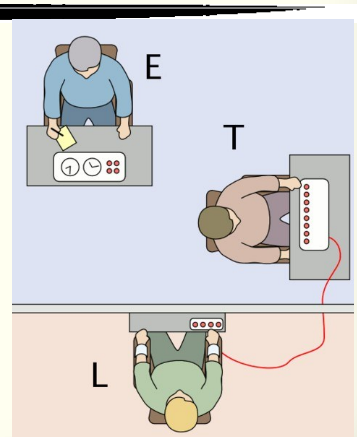
- Left skew
- Points bend down and to the right of the line.



- Long tails (wider than the normal distribution)
- Points start below the line, bend to follow it, and end above it.

The Milgram Experiment

- Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.



The Milgram Experiment

- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.

- Milgram found that about 65% of people would obey authority and give such shocks.

- Over the years, additional research suggested this number is approximately consistent across communities and time.

$$P(\text{shock}) = 0.65$$

Bernoulli Random Variable

- Each person in Milgram's experiment can be thought of as a trial.

- A person is labeled a success if she refuses to administer a severe shock, and failure if she administers such shock.

- Since only 35% of people refused to administer a shock, probability of success is p = 0.35

- When an individual trial has only two possible outcomes, it is called a Bernoulli random variable.

Geometric Distribution

- Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{\text{st}} \text{ person refuses}) = 0.35$$

- ... the third person?

$$P(1^{\text{st}} \text{ and } 2^{\text{nd}} \text{ shock; } 3^{\text{rd}} \text{ refuses}) = (0.65) \times (0.65) \times (0.35) = 0.15$$

- ... the tenth person?

$$P(9 \text{ shock; } 10^{\text{th}} \text{ refuses}) = \underbrace{(0.65) \times \cdots \times (0.65)}_{\text{9 times}} \times (0.35) = 0.15$$

Geometric Distribution

- Geometric distribution describes the waiting time until a success for independent and identically distributed (iid) Bernoulli random variables.
  - independence: outcomes of trials don't affect each other
  - identical: the probability of success is the same for each trial
- If p represents probability of success, $(1-p)$ represents probability of failure, and n represents number of independent trials

$$P(\text{success on the } n^{\text{th}} \text{ trial}) = (1-p)^{n-1}p$$

- Mean and standard deviation of geometric distribution:

$$\mu = \frac{1}{p} \quad , \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

Binomial Distribution

- Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?
- Let's call these people (A), (B), (C), and (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock

$$\text{Scenario 1: } \underset{\text{(A) } \textit{refuse}}{0.35} \times \underset{\text{(B) shock}}{0.65} \times \underset{\text{(C) shock}}{0.65} \times \underset{\text{(D) shock}}{0.65} = 0.0961$$

OR

$$\text{Scenario 2: } \underset{\text{(A) shock}}{0.65} \times \underset{\text{(B) } \textit{refuse}}{0.35} \times \underset{\text{(C) shock}}{0.65} \times \underset{\text{(D) shock}}{0.65} = 0.0961$$

OR

$$\text{Scenario 3: } \underset{\text{(A) shock}}{0.65} \times \underset{\text{(B) shock}}{0.65} \times \underset{\text{(C) } \textit{refuse}}{0.35} \times \underset{\text{(D) shock}}{0.65} = 0.0961$$

OR

$$\text{Scenario 4: } \underset{\text{(A) shock}}{0.65} \times \underset{\text{(B) shock}}{0.65} \times \underset{\text{(C) shock}}{0.65} \times \underset{\text{(D) } \textit{refuse}}{0.35} = 0.0961$$

- The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities:

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial Distribution

- The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as:

$$\#\ of\ scenarios \times P(single\ scenario)$$

# of scenarios: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

R
```
> choose(n,k)
```

$P(single\ scenario) = p^k(1-p)^{n-k}$

probability of success to the power of number of successes

probability of failure to the power of number of failures

- The Binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p.

Binomial Conditions

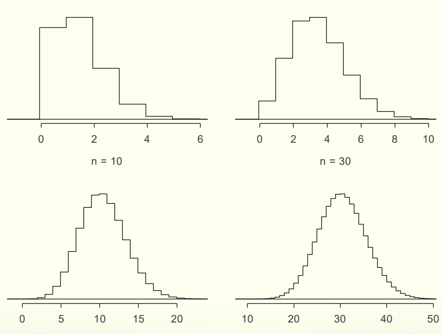
1. The trials must be independent
2. The number of trials, n, must be fixed
3. Each trial outcome must be classified as a success or a failure
4. The probability of success, p, must be the same for each trial

R
```
> dbinom(8, size = 10, p = 0.13)
[1] 2.77842e-06
```

- Expected value (mean) of binomial distribution: $\mu = np$
- Standard deviation of binomial distribution: $\sigma = \sqrt{np(1-p)}$

Normal Approximation to Binomial

- Hollow histograms of samples from the binomial model where $p = 0.1$ and n = 10, 30, 100, and 300. What happens as n increases?



Normal Approximation to Binomial

- When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.
- $X \sim Binomial(n, p)$, $Z \sim N(0,1)$

$$P(k_1 \leq X \leq k_2) \approx P(\frac{k_1 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{k_2 - np}{\sqrt{np(1-p)}})$$

- How large is large enough?
- The sample size is considered large enough if the expected number of successes and failures are both at least 10:

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

Example

A Comparison of Common Users across Instagram and Ask.fm to Better Understand Cyberbullying

Homa Hosseinmardi
Shaosong Li, Zhili Yang and Qin Lv
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado, USA
Email: homa.hosseinmardi@colorado.edu

Rahat Ibn Rafiq
Richard Han and Shivakant Mishra
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado, USA
Email: rahat.rafiq@colorado.edu

- 39% of Instagram accounts are private.
- Other findings:
  - Average Instagram user has 150 followers.
  - Only 30% of Iranian profiles in Instagram is set to private.
- P(average Instagram user in Iran has 60 or more private followers) = ?

Example

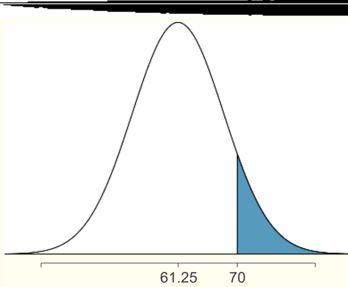
- We are given a binomial distribution with $n = 150$ and $p = 0.3$
  1) n = 150, fixed
  2) each trial has 2 outcomes: private / public
  3) p = 0.3 and fixed in all trials
  4) Independence
- $P(X \geq 60) = ?$

$$P(X \geq 60) = P(X = 60) + P(X = 61) + \cdots + P(X = 150)$$

$$\mu = np = 150 \times 0.3 = 45$$
$$\sigma = \sqrt{150 \times 0.3 \times 0.7} = 5.61$$

$$P(X \geq 60) = P\left(Z \geq \frac{60 - 45}{5.61}\right) = P(Z > 2.67)$$

Example



	Second decimal place of Z				
Z	0.03	0.04	0.05	0.06	0.07
2.4	0.9924	0.9927	0.9929	0.9931	0.9932
2.5	0.9943	0.9945	0.9946	0.9948	0.9949
2.6	0.9957	0.9958	0.9960	0.9961	0.9962

Approximation: $P(X \geq 60) = P(Z > 2.67) = 1 - P(Z < 2.67) = 1 - 0.9962 = 0.0038$

Exact Value:

R
```
> sum(dbinom(60:150, size=150, p=0.3))
[1] 0.0057
```

Continuity Correction Factor

- The binomial distribution is a discrete distribution while the normal distribution is a continuous distribution.
- The basic difference here is that with discrete values, we are talking about heights but no widths, but with the continuous distribution we are talking about both heights and widths.
- The correction is to either add or subtract 0.5 of a unit from each discrete X-value.

Example:

Discrete	Continuous
$X = 6$	$5.5 < X < 6.5$
$X > 6$	$X > 6.5$
$X \geq 6$	$X > 5.5$
$X < 6$	$X < 5.5$
$X \leq 6$	$X < 6.5$

Continuity Correction Factor

$$P(k_1 \leq X \leq k_2) \approx P(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}})$$

Therefore:

$$P(X \geq 60) = P\left(Z > \frac{59.5 - 45}{5.61}\right) = P(Z > 2.57)$$

$$= 1 - P(Z < 2.57) = 1 - 0.9949 = 0.0051$$

Poisson Distribution

- The Poisson distribution is used to model the number of events occurring within a given time interval. The formula for the Poisson pmf is:

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

- λ is the shape parameter which indicates the average number of events in the given time interval.

Siméon Poisson (1781–1840)

- The distribution was derived by the French mathematician Siméon Poisson in 1837, and the first application was the description of the number of deaths by horse kicking in the Prussian army!

Applications of Poisson Distribution

- Some events are rather rare: they don't happen that often.
  - For instance, car accidents are the exception rather than the rule.
  - Still, over a period of time, we can say something about the nature of rare events.
- The Poisson distribution can be a useful model for such events.
- Other phenomena that often follow a Poisson distribution are death of infants, the number of misprints in a book, the number of customers arriving, and …

R

```
> dpois(x = 1, lambda = 2.5)
[1] 0.2052125
```

Properties of Poisson Distribution

- The mean and variance are both equal to λ.
- The sum of independent Poisson variables is a further Poisson variable with mean equal to the sum of the individual means.
- As well as cropping up in the situations already mentioned, the Poisson distribution provides an approximation for the Binomial distribution.
- If n is large and p is small, then the Binomial distribution with parameters n and p, is well approximated by the Poisson distribution with parameter $\lambda = np$, i.e. by the Poisson distribution with the same mean.

Example

- X has Binomial distribution with $n = 100$, $p = 0.075$. Calculate the probability of fewer than 10 successes: $P\{X < 10\}$.

R
```
> sum(dbinom(0:9, size=100, p=0.075))
[1] 0.7832687
```

- The Poisson approximation to the Binomial states that λ will be equal to np, i.e. $\lambda = 100 \times 0.075 = 7.5$

R
```
> sum(dpois(0:9, lambda=7.5))
[1] 0.7764076
```

OR

R
```
> ppois(9, lambda=7.5)
[1] 0.7764076
```

Distributions in R

dnorm(x, mean = 0, sd = 1) : density function of a normal dist.
pnorm(q, mean = 0, sd = 1) : distribution function (CDF) of normal dist.
qnorm(p, mean = 0, sd = 1) : quantile function of normal dist.
rnorm(n, mean = 0, sd = 1) : generate n random numbers from the normal dist.

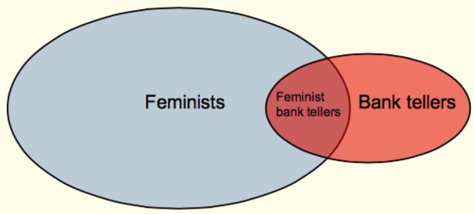
- We have similar functions for other distributions:
  - dpois, ppois, qpois, rpois
  - dbinom, pbinom, qbinom, rbinom
  - …

The Linda Experiment

- **Experiment by Tversky and Kahneman:** Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Rank the following scenarios:
  - Linda is a teacher in elementary school.
  - Linda works in a bookstore and takes yoga classes.
  - Linda is active in the feminist movement.
  - Linda is a psychiatric social worker.
  - Linda is a member of the League of Women Voters.
  - Linda is a bank teller.
  - Linda is an insurance salesperson and takes yoga classes.
  - Linda is a bank teller and is active in the feminist movement.

Conjunction Fallacy

- Even if you assign a very high probability that Linda is a feminist, and a very low probability for being a bank teller, the *combined* probability of both cannot be higher than the probability that Linda is a bank teller.



$$P(A \cap B) \leq P(A)$$

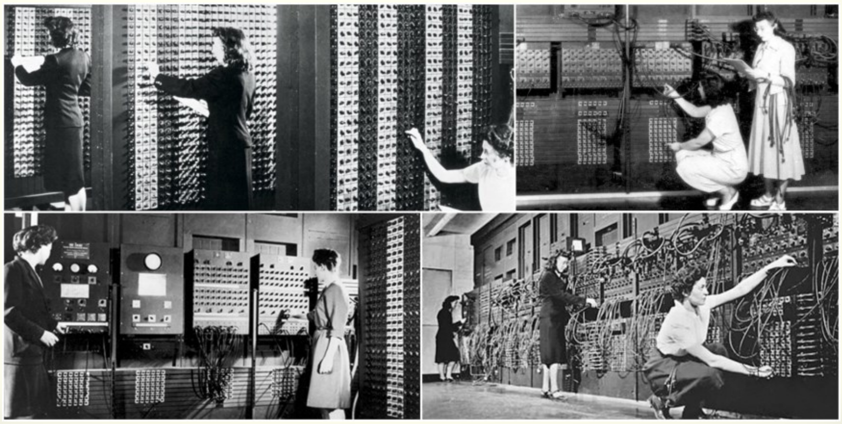
$$P(A \cap B) \leq P(B)$$

- Following the same intuition, Linda may resemble a feminist bank teller more than she resembles a bank teller, but it cannot be more likely. As a result, people tend to make some very plain statistical mistakes.

Monte Carlo Simulation

- Ulam, recovering from an illness, was playing a lot of solitaire.
- Tried to figure out probability of winning, and failed.
- Thought about playing lots of hands and counting number of wins, but decided it would take years.
- Asked Von Neumann if he could build a program to simulate many hands on ENIAC

ENIAC



Monte Carlo Simulation

- A method of estimating the value of an unknown quantity using the principles of inferential statistics.
- Inferential statistics:
  - Population: a set of examples
  - Sample: a proper subset of a population
  - Key fact: a random sample tends to exhibit the same properties as the population from which it is drawn

Interview Question

- You have two gift cards from your favorite coffee shop. Each gift card is loaded with 50 free coffees. The cards are identical, so it is hard to tell them apart. Every time, you randomly pick one of the cards and use it to pay for your free drink. One day the barista tells you that she can't accept the card as it doesn't have any drinks left.

- What is the mean number of free drinks on the other card?


```python
import random

def monte_carlo_simulation(num_simulations, num_initial_coffees):

    total_remaining_coffees = 0

    for _ in range(num_simulations):
        coffees_on_first_card = num_initial_coffees
        coffees_on_second_card = num_initial_coffees

        while coffees_on_first_card > 0 and coffees_on_second_card > 0:
            # Randomly select one of the cards
            chosen_card = random.choice([1, 2])

            # Deduct a coffee from the chosen card
            if chosen_card == 1:
                coffees_on_first_card -= 1
            else:
                coffees_on_second_card -= 1

        # Add the remaining coffees on the other card
        total_remaining_coffees += max(coffees_on_first_card, coffees_on_second_card)

    # Calculate the mean number of remaining coffees on the other card
    mean_remaining_coffees = total_remaining_coffees / num_simulations

    return mean_remaining_coffees
```

Monte Carlo Simulation

```python
# Parameters
num_simulations = 100000
num_initial_coffees = 50

# Perform Monte Carlo simulation
mean_remaining_coffees = monte_carlo_simulation(num_simulations, num_initial_coffees)
print(f"Mean number of free drinks on the other card: {mean_remaining_coffees}")
```

```
Mean number of free drinks on the other card: 7.95291
```