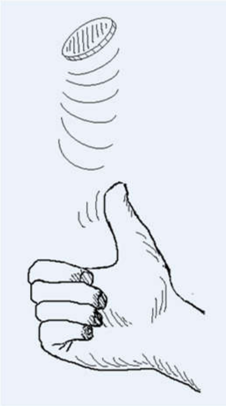
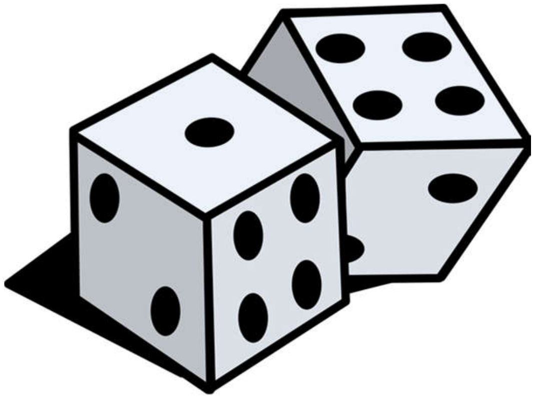
Introduction to Data Science

A Review of Probability Theory

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.



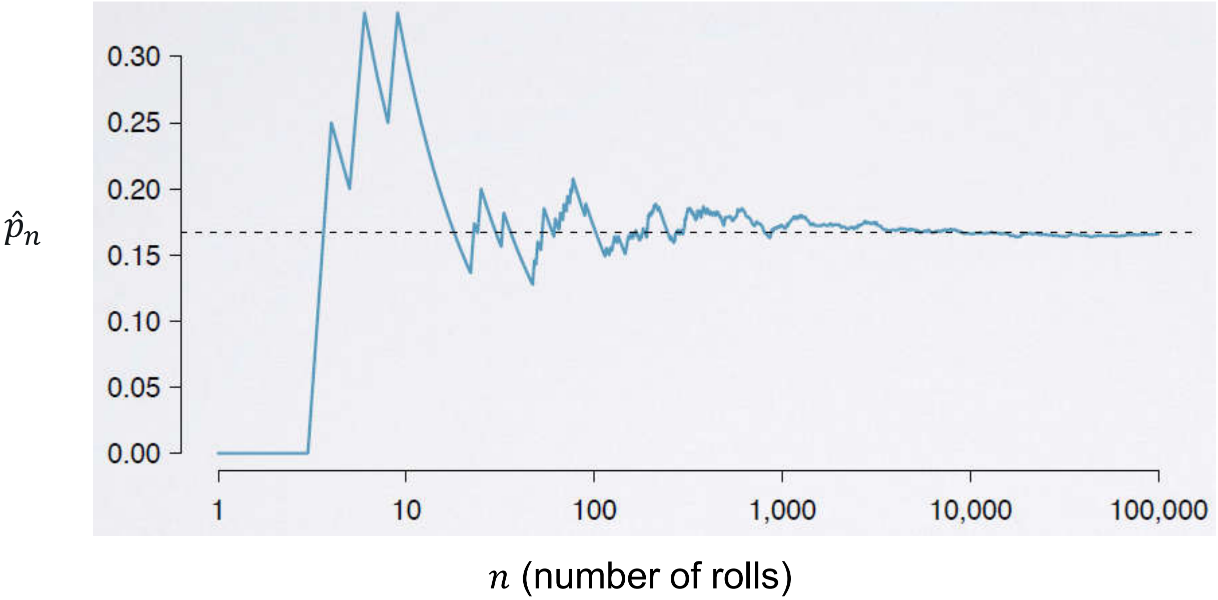


Probability

- 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).

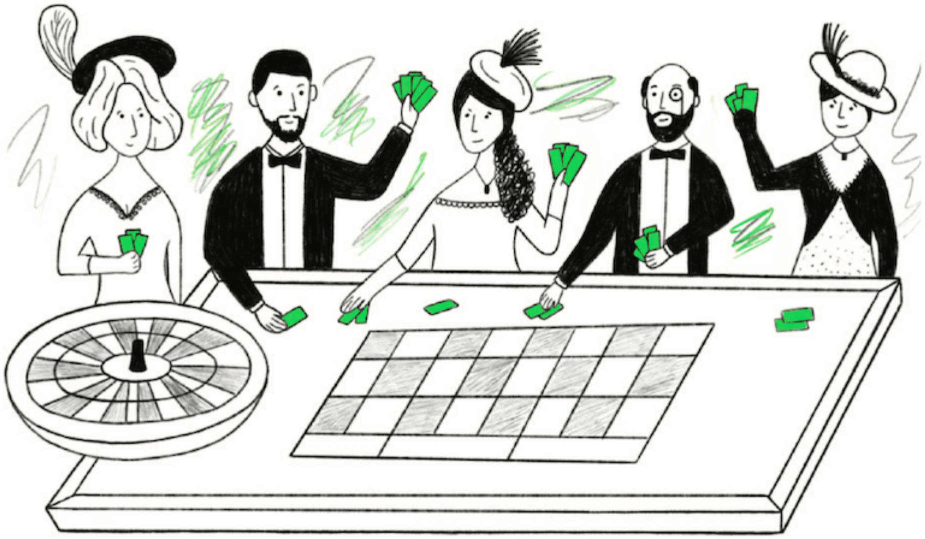


Gambler's Fallacy

- 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

- The mistaken belief that because something has happened more frequently than usual, it’s now less likely to happen in future.



Gambler’s Fallacy

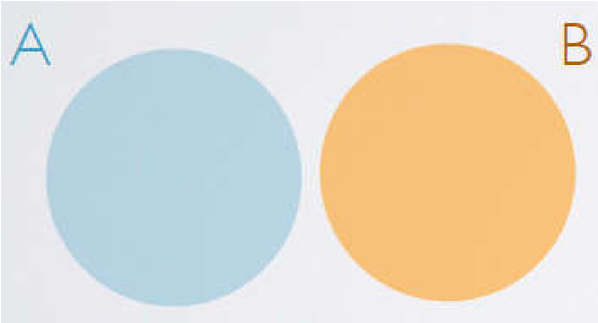
- On August 18, 1913, at a casino in Monte Carlo, black came up a record twenty-six times in succession [in roulette] … There was a near panicky rush to bet on red, beginning about the time black had come up a phenomenal fifteen times.

  -- Huff and Geis, How to Take a Chance

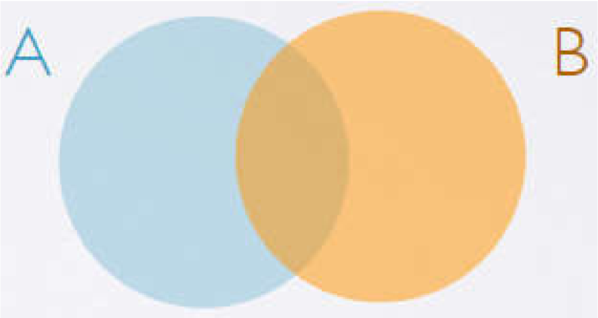
- Probability of 26 consecutive reds
  - 1 in 66.6 million
- Probability of 26 consecutive reds when previous 25 rolls were red
  - Almost 1/2

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.
- 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.



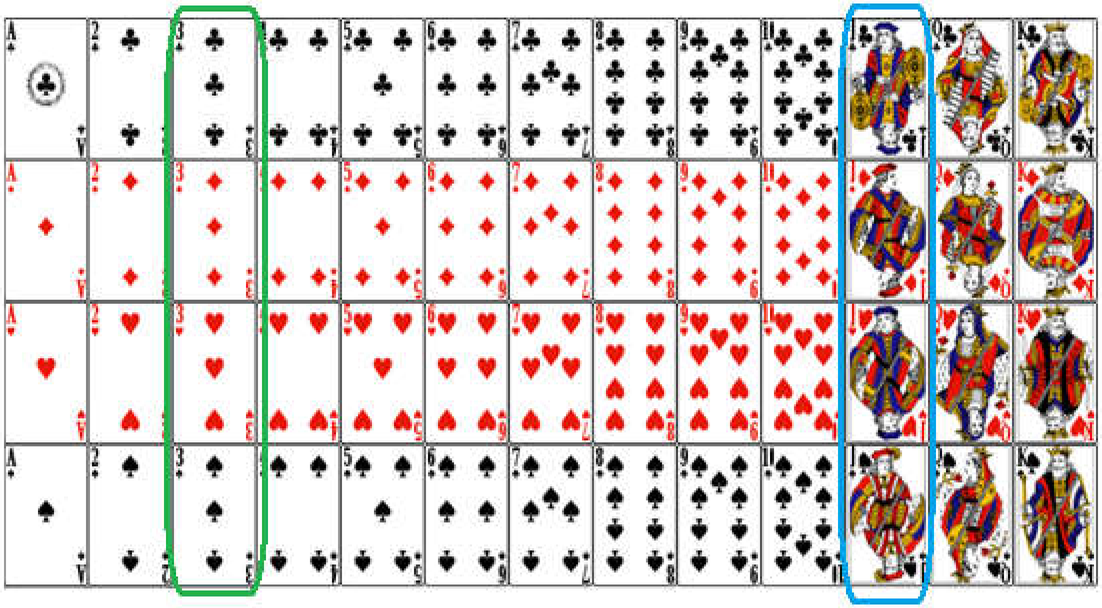
$P(A \cap B) = 0$



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

Union of Disjoint Events

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



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

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

$4/52 + 4/52 =$

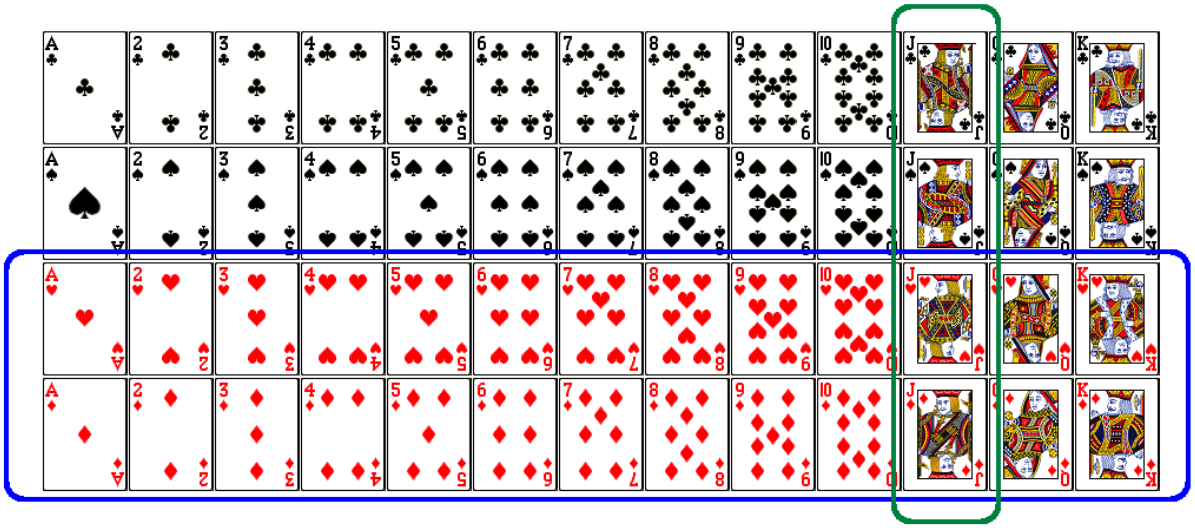
≈ 0.154

- 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

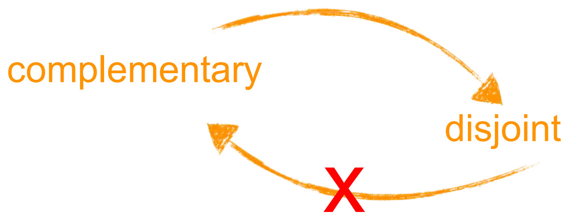
- Complementary events are two mutually exclusive events whose probabilities add up to 1.

complementary

one toss	H	T
probability	0.5	0.5

complementary

two tosses	HH	HT	TH	TT
probability	0.25	0.25	0.25	0.25



complementary

disjoint

X

Independence

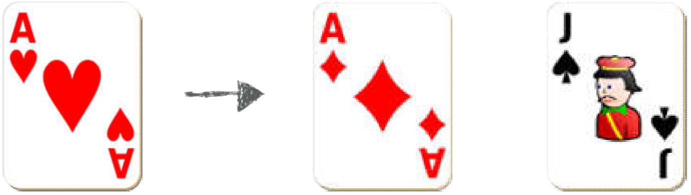
- Two random processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

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

A Survey

- Between January 9-12, 2013, SurveyUSA interviewed a random sample of 500 North Carolina 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.

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

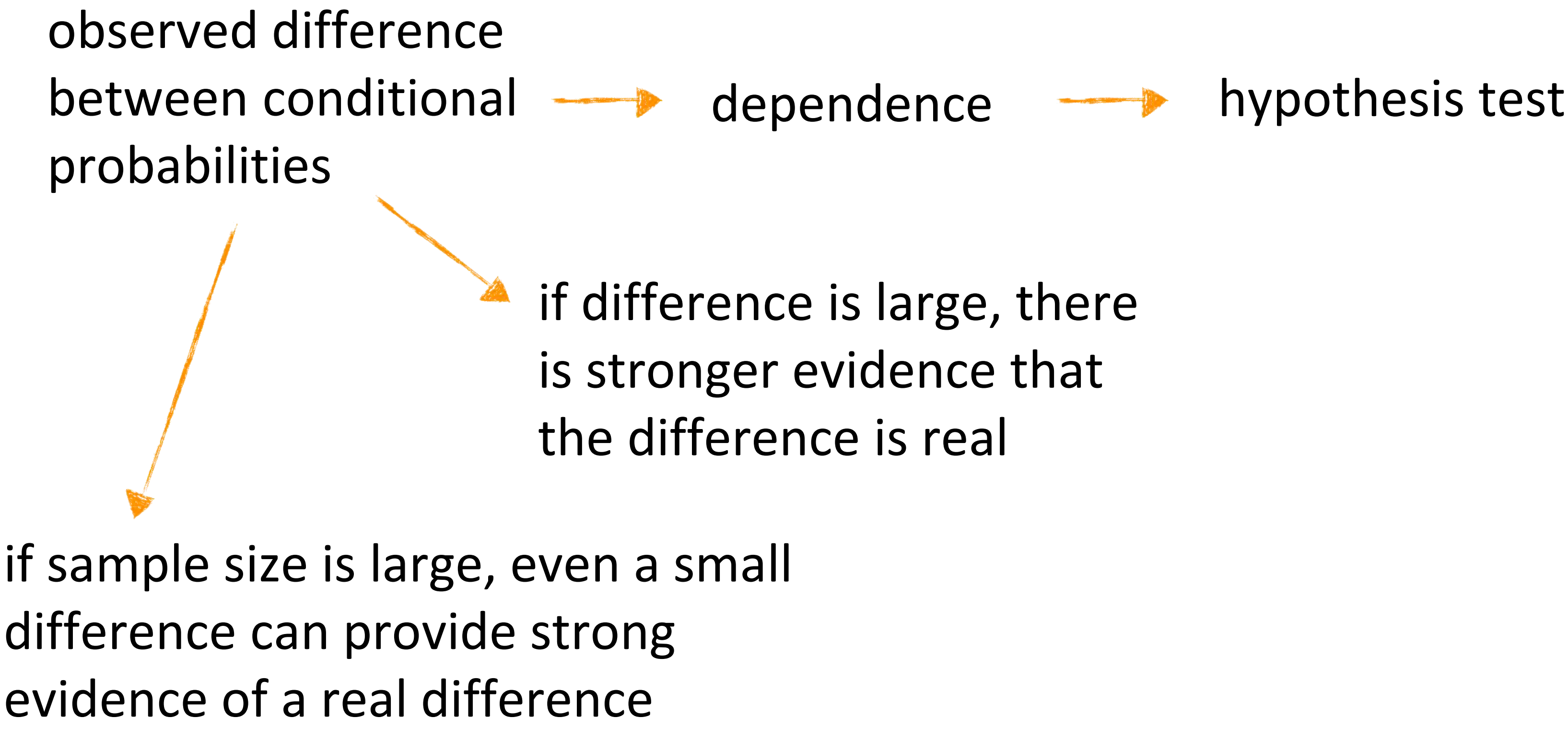
  P(protects citizens | White) = 0.67

  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

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 $A_1, A_2, \ldots, A_k$ are independent:

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

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:

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

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".

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 $A_1, \ldots, A_n$:

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

- 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:

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

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

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 ● , 3 ● , 2 ●

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

- Suppose you 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} draw: 5 ● , 3 ● , 2 ●

2^{nd} draw: 5 ● , 3 ● , 2 ●

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

Sampling with Replacement

- Suppose you 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?

1st draw: 5 ● , 3 ● , 2 ●

2nd draw: 5 ● , 3 ● , 2 ●

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

- When drawing with replacement, probability of the second chip being blue does not depend on the color of the first chip, 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

- 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?

1st draw: 5 🔴 , 3 🔵 , 2 🟡

2nd draw: 5 🔴 , 2 🔵 , 2 🟡

$Prob(2^{nd}$ chip $B | 1^{st}$ 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]?

$$E[Y] = 5\ (5/165) + 10\ (10/165) + 150\ (150/165) = 22635/165 \approx 137$$

- Note: E[Y] is students' perception of class size.

Variance

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

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

- 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

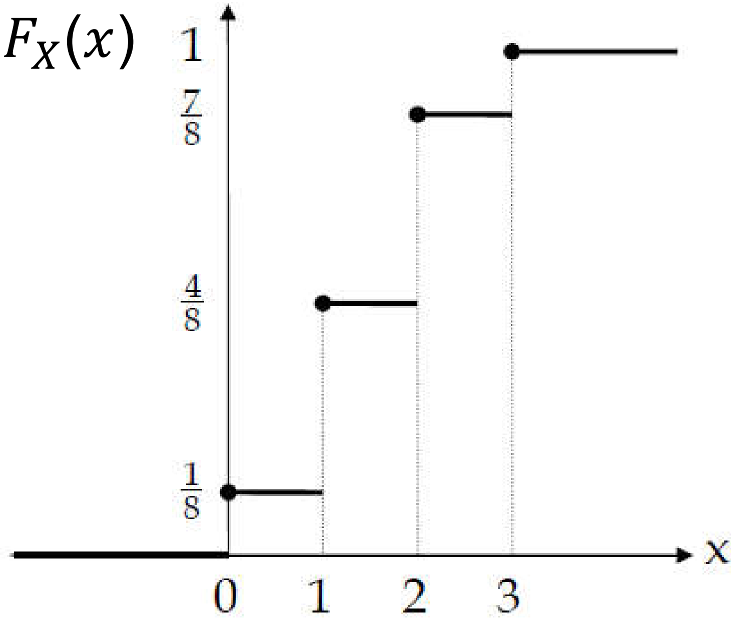
- Unlike a discrete random variable, the values that a continuous random variable X can take is uncountable, so we usually have:

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

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

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

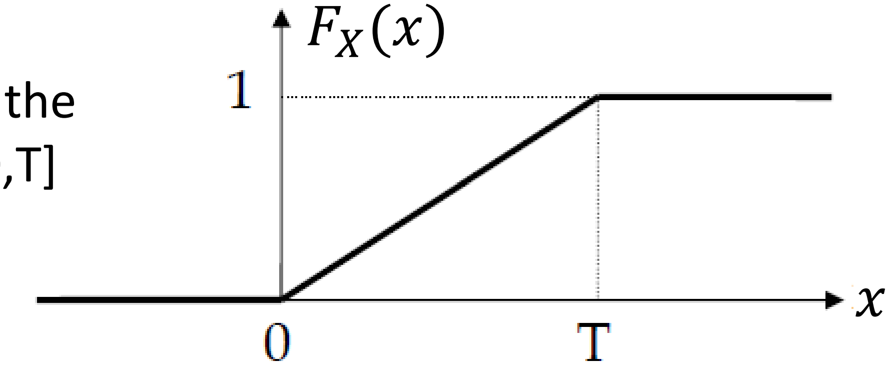
- For the 3 coins example we have:



Continuous Distribution

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

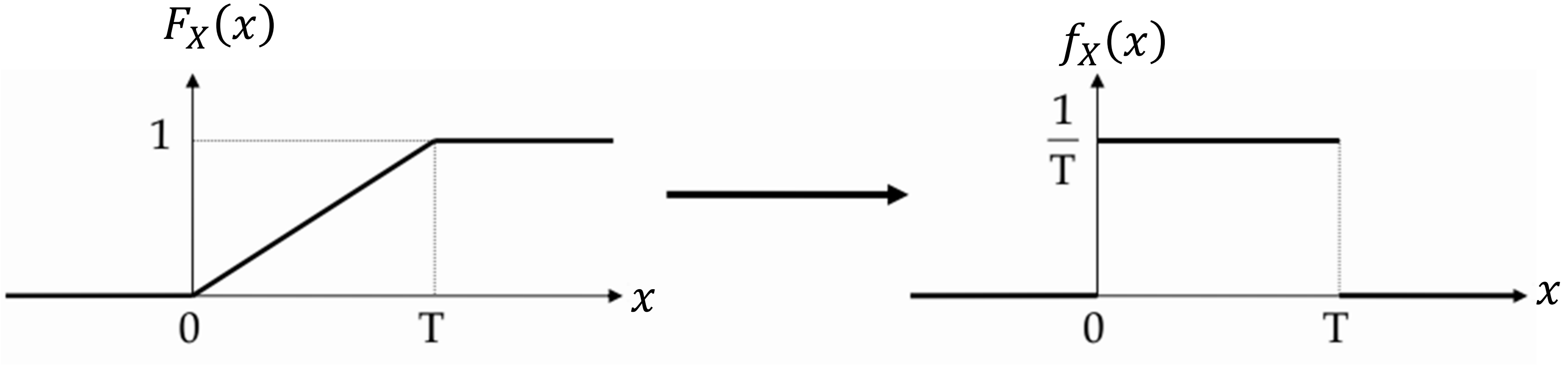
X = random variable that represents the time of a phone call in the interval [0,T]



- 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

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



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

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:

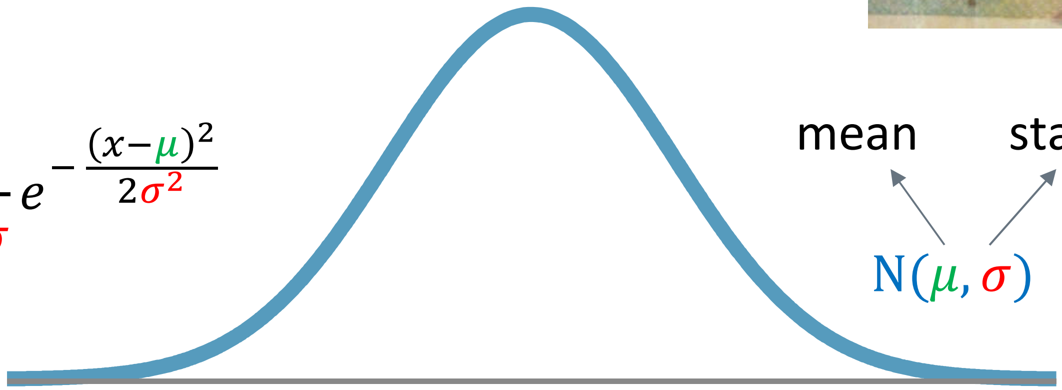
$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2\, f_X(x) dx$$

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

Normal (Gaussian) Distribution

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

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



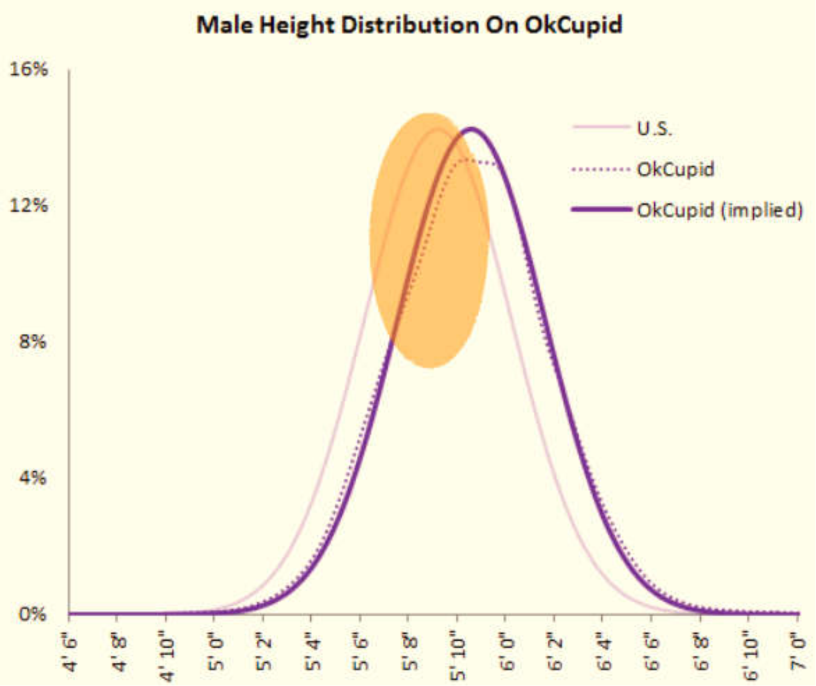
mean

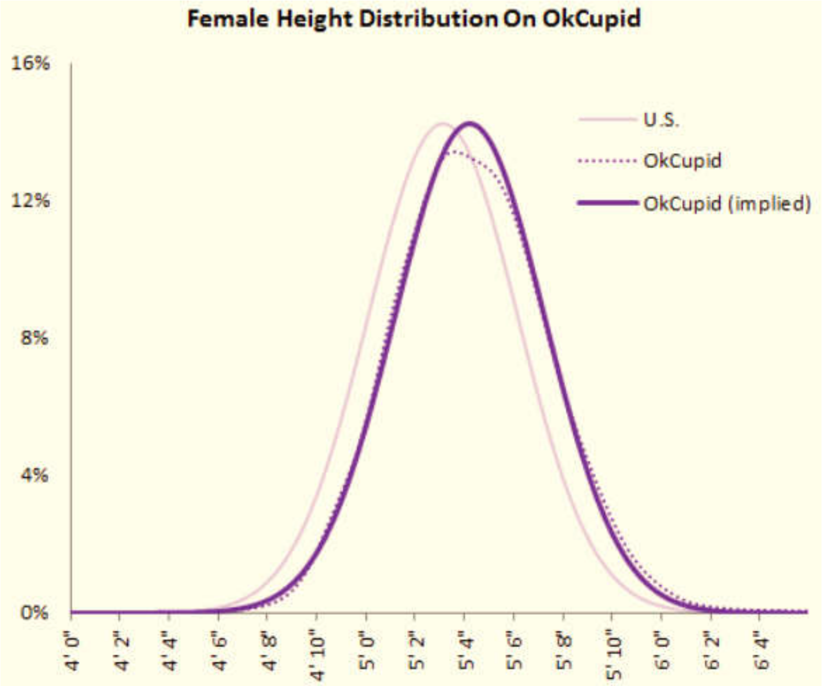
standard deviation

$N(\mu, \sigma)$

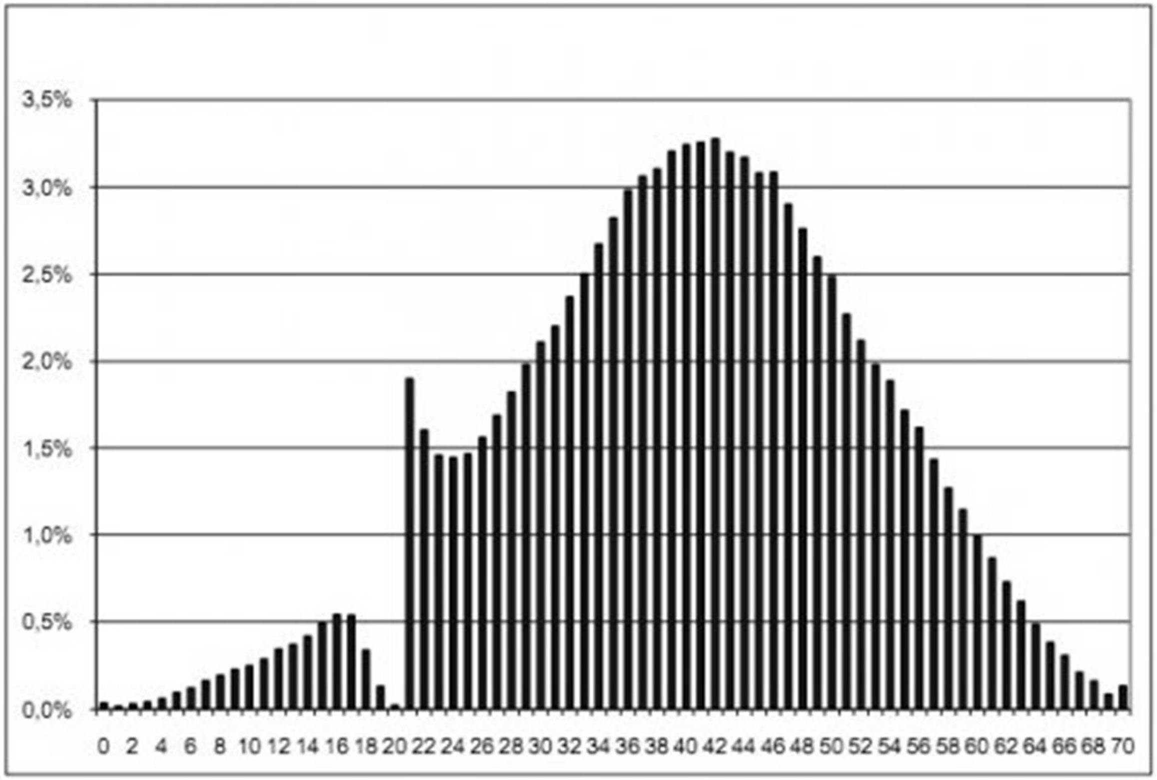
Normal Distribution in Real Life

- Many random variables in nature has normal distribution.

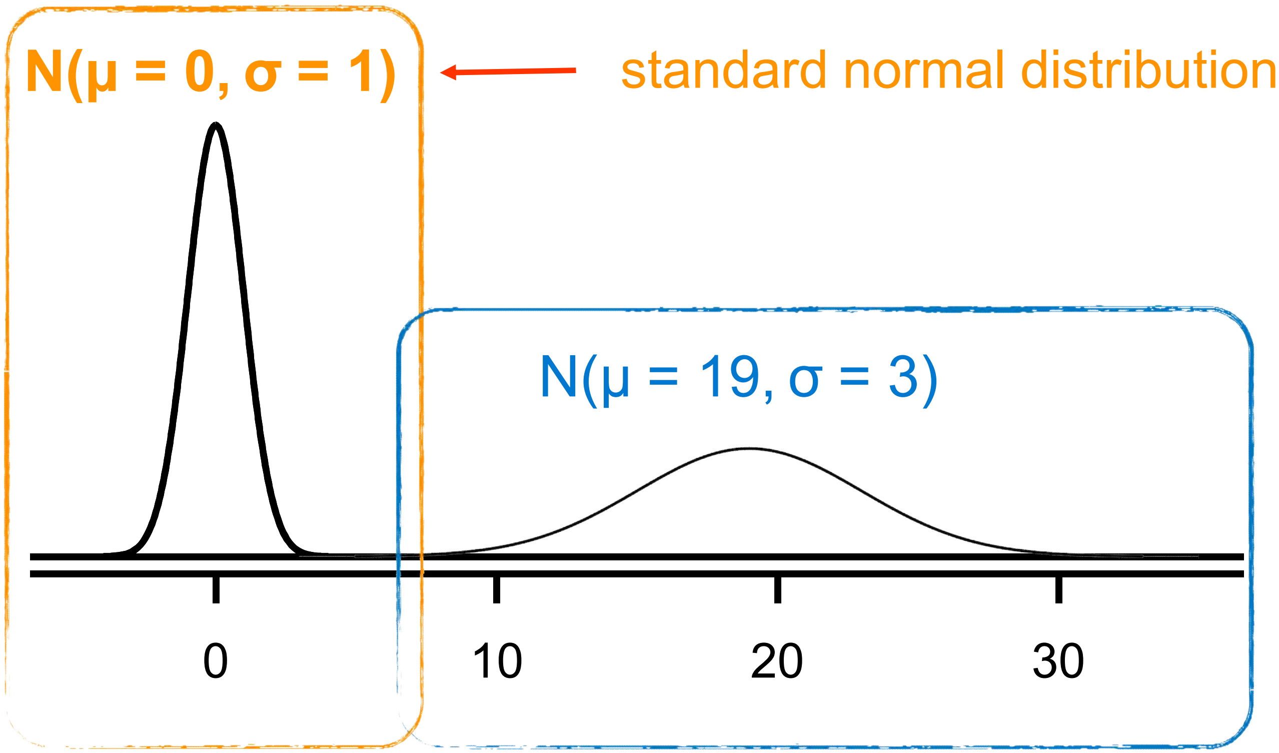




Final Grade Distribution in Poland

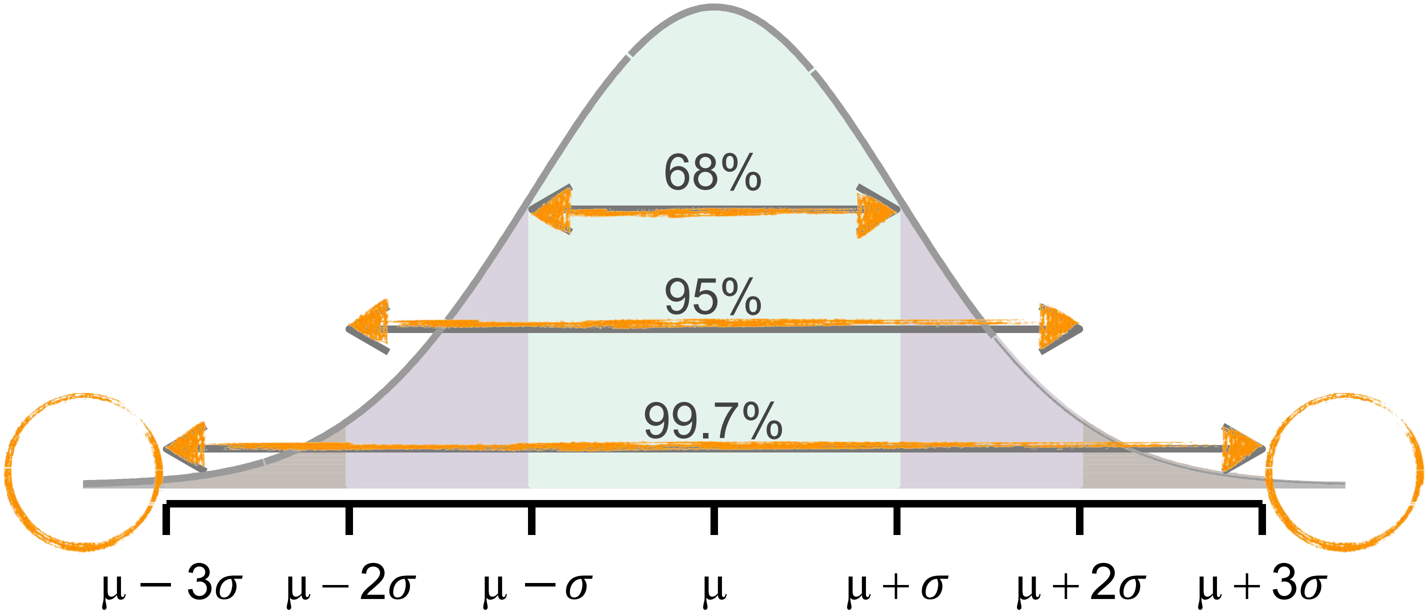


Normal Distribution



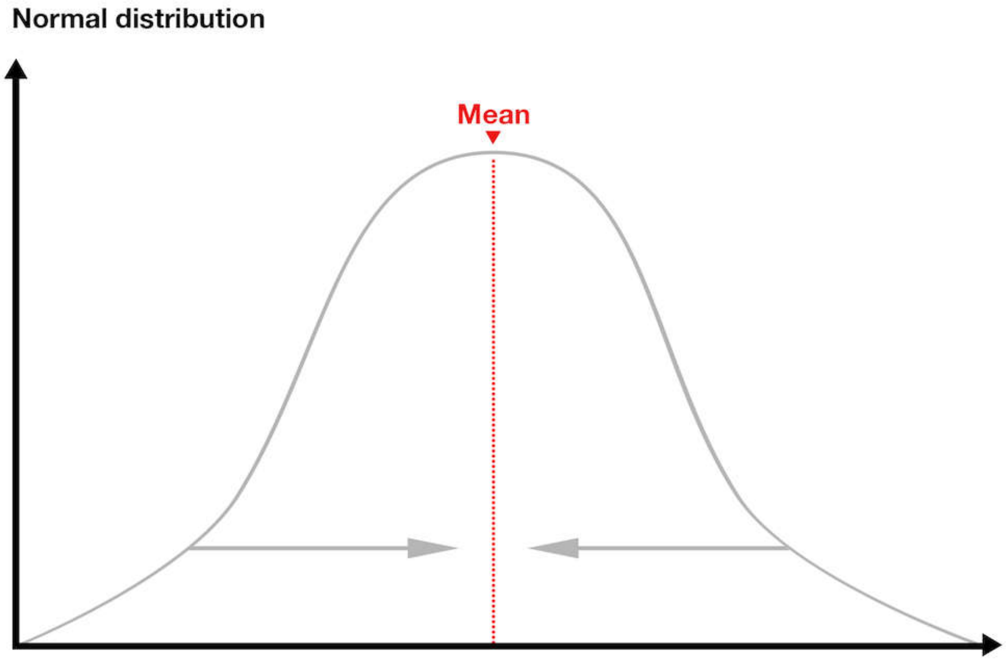
standard normal distribution

Empirical Rule (68 - 95 - 99.7% rule)



Regression to the Mean

Why perfection rarely lasts?

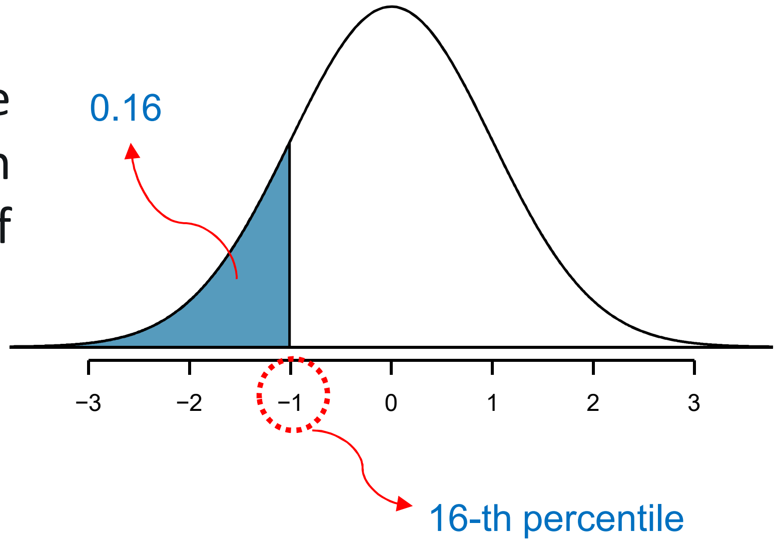


Regression to the Mean

- 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”!

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.



- 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.

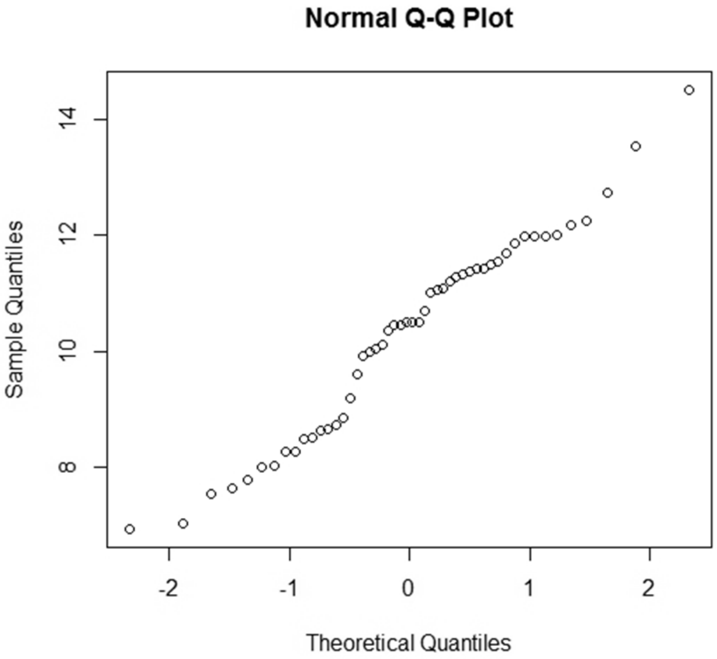
Quantile

- **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

- 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.

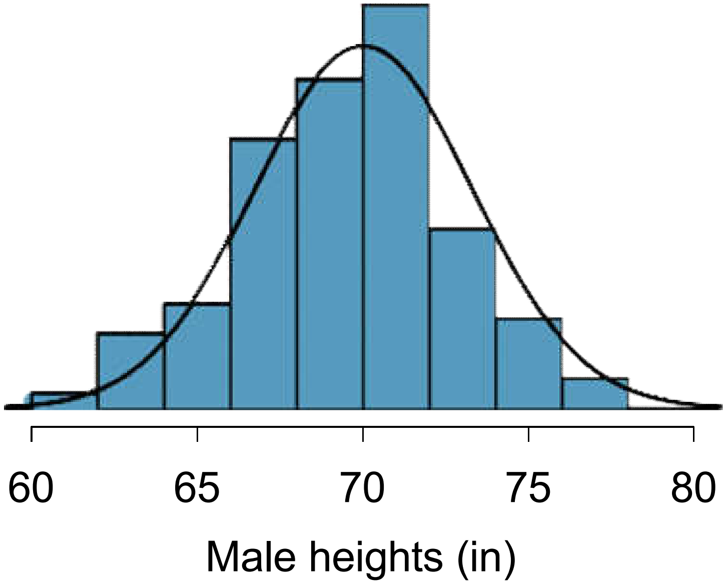
Q-Q Plots

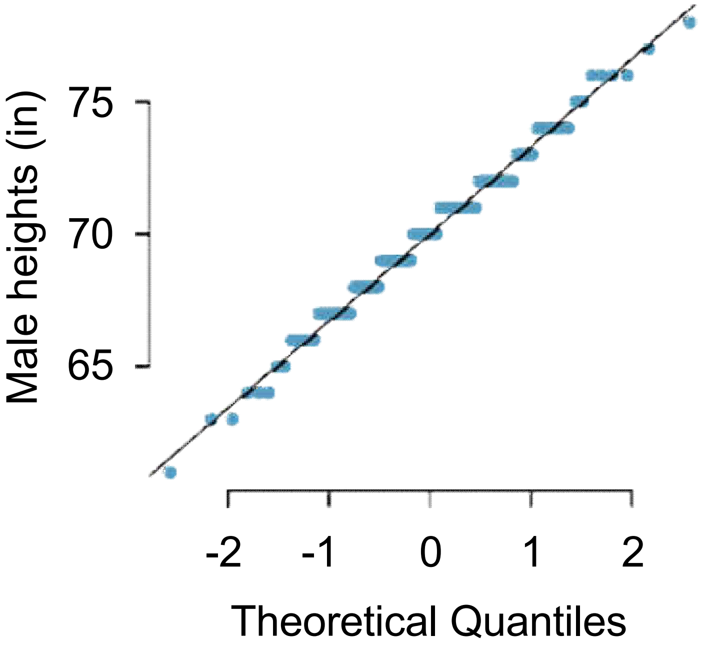
- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.
  - It's just a visual check, not an air-tight proof.



Normal Probability Plot

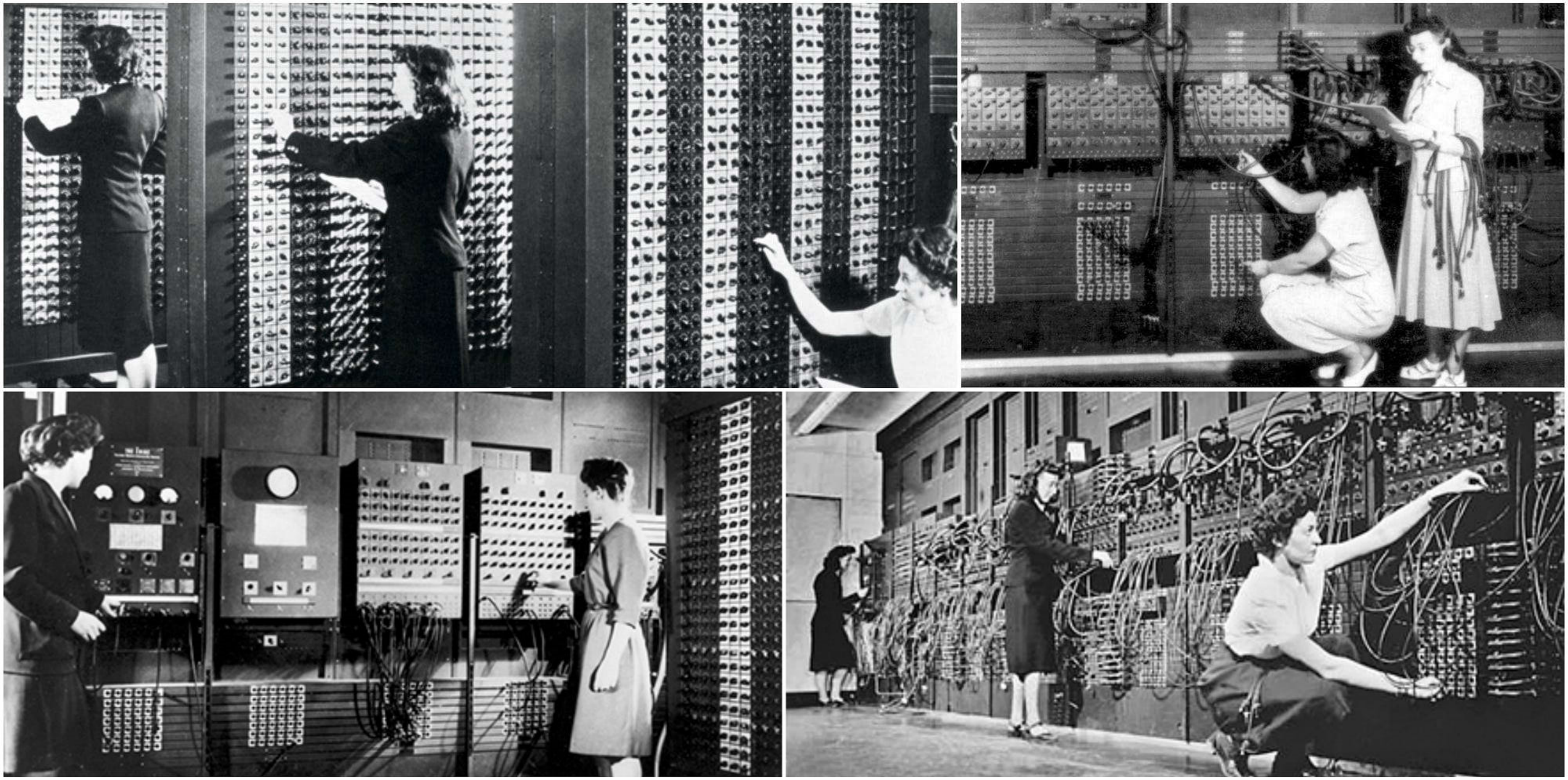
- A histogram and normal probability plot of a sample of 100 male heights.





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.



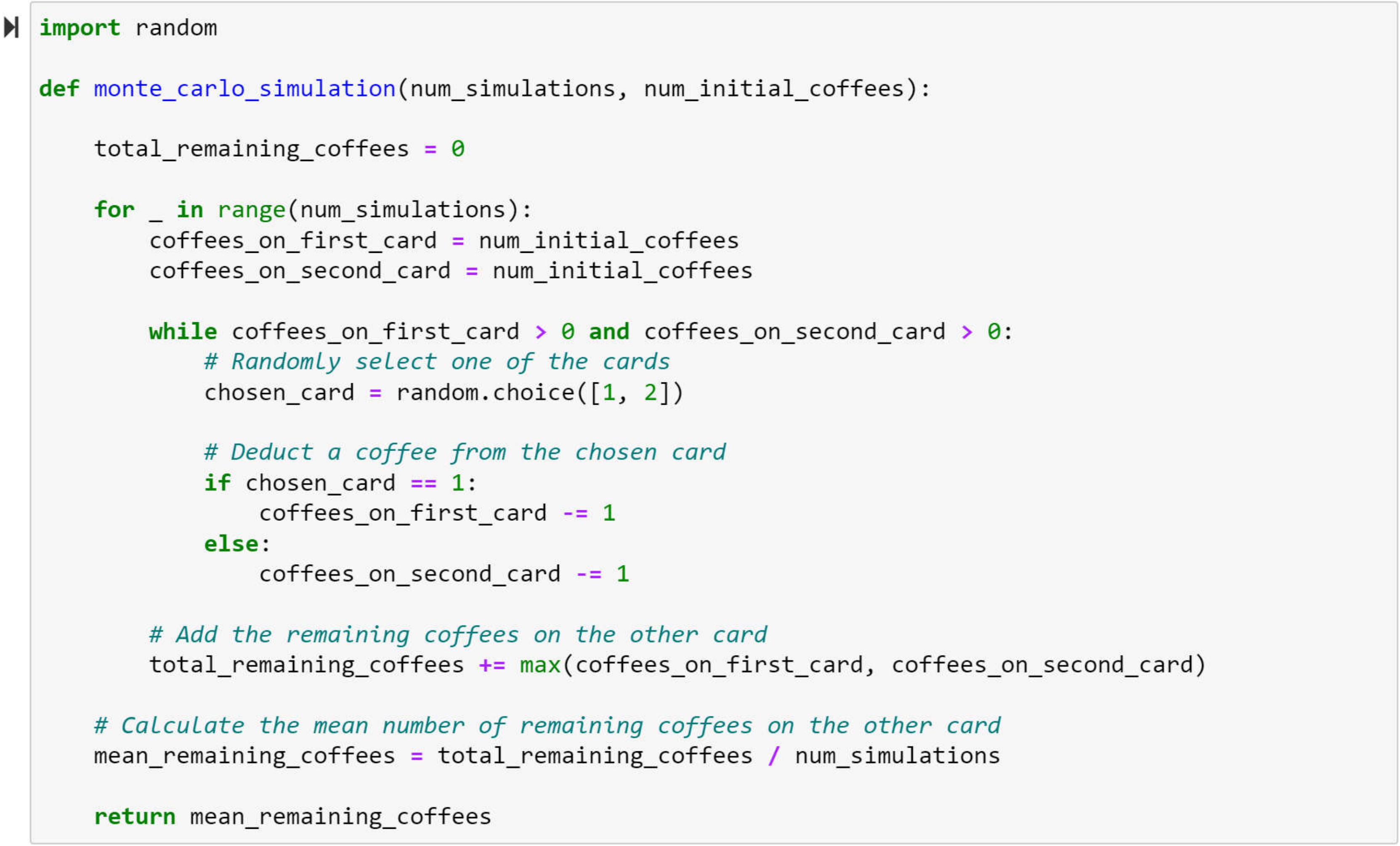
Monte Carlo Simulation

- 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

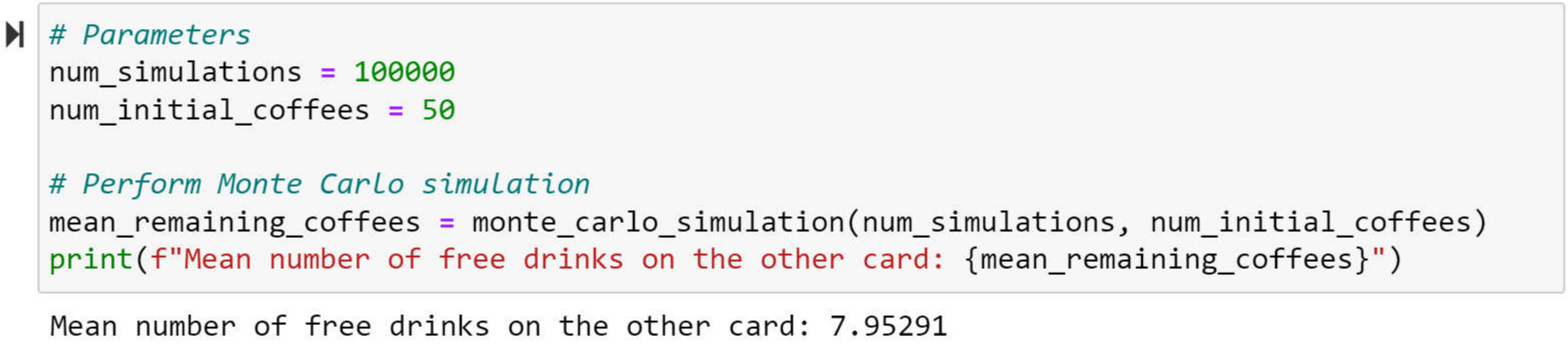
Interview Question

- 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



```
# 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
```