BIOSTAT 701

Introduction to Statistical Theory and Methods I

Lynn Lin

Preparation

- **Some basics about RV**: https://duke.zoom.us/rec/play/HZwTlXZFFnXEaBOLcVLBRIMkXMWUScrCst72WOkYRV2RhYwMMxO8CYPuKnADc0-M5VYpaN_3DMTSWs1O.AMP3HZua4Ykn3iSt?canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fduke.zoom.us%2Frec%2Fshare%2Ff6GP-fBykj8hFJW07zMUFNayWxGSQVbjqHrE_spq4RA5h5vHQXhv153MHnW7IZZu.tNpkICAmZPgHSKaY

Preparation

- **Joint and conditional probability**: https://duke.zoom.us/rec/share/MJq1llADHKIj9EO1yBOrwMMcisIAwCCes1_zOYDTXAhO-K1E2yfciiGsLpVllbT-.K74zBnD3i2v5Ni4B?startTime=1645213554000
- **Independence**: https://duke.zoom.us/rec/share/mCuhuuqFd31tsng7E_9dxkYoILxR5f4zVEkNwW9VQJQ_zPg-JZOOI23rgDmMSm0z.5qOR2kIyu34_nXSu?startTime=1646155389000

Nature of randomness

- People used to think everything can be predicted.
- We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes. *–Pierre Simon Laplace, A Philosophical Essay on Probabilities*

Nature of randomness

- Scientific evidence of randomness: Heisenberg’s uncertainty principle.
- Randomness is important in many respects: Gambling, Government Planning, Finance, Biology,…

Why study statistics?

- To evaluate printed numerical facts: Your company's annual report printed that the sales next year are expected to be $11.50 million with a standard deviation of $1.2 million.

- To interpret the results of sampling or to perform statistical analysis in your work: You are asked to project your company's sale for next year.

- To make inference about the population using information collected from the sample: A Readers digest/Gallup survey on the drinking habits of Americans estimated the percentage of adults across the country who drink beer, wine, or hard liquor, at least occasionally.

Data challenge

- Due to various constraints, for example, time or budget, one can only sample from the population instead of take a census of the population. We need a sample that closely represents the population. One way is to obtain a random sample.
- What do statistician do?

representative data

What do statistician do?

- Gathering data: How do we collect the data? Figure
- Summarizing data: How do we present the data?
- Analyzing data: What method should we use to analyze the data?

link.

- Drawing conclusions and reporting the results of the analyses

Example

- You want to know the average amount of time American people spend browsing the web. How are you going to perform the above four steps?

age group.

1. collect data: range of different age.
2. present: table / dataset.
3. analyze.
4. conclude.

provider.

apple: screen time.

How many?

(sample size

One answer will be:

- Take a random sample of American people. How many to sample depends on how accurate you want your inference to be and the margin of error you can tolerate.

- Summarize data from the sample.

- Analyze the data to obtain confidence interval.

- Use written report with graphical, tabular and numerical displays.

How to collect data for surveys?

- Personal interview: People usually respond when interviewed by a person but their answers may be influenced by the interviewer.
- Telephone interview: Cost effective but need to keep it short since respondents tend to be impatient.
- Self-administered questionnaires: Cost effective but the response rate is lower and the respondents may be a biased sample.
- Direct observation: For certain quantities of interest, one may be able to measure it from the sample.
- Web-based survey,..., etc.

Principles of experimental design

- The following principles of experimental design has to be followed to enable a researcher to conclude that differences in the results of an experiment not reasonably attributable to chance are likely to have been caused by the treatments.
    - Control: Need to control for effects due to factors other than the ones of primary interest.

      avoid untensional bias cage...?
    - Randomization: Subjects should be randomly divided into groups to avoid unintentional selection bias in the groups.
    - Replication: A sufficient number of subjects should be used to ensure that randomization creates groups that resemble each other closely and to increase the chances of detecting differences among the treatments when such differences actually exist.

Terminology

- A (stochastic) experiment is a procedure that yields one of a given set of possible outcomes
- We deal with experiments whose results aren't fully predictable. Those results are described through random variables. We sometimes call the process by which random variables are created the data generating mechanism.
- The sample space S of the experiment is the set of possible outcomes
- An event is a subset of sample space
- A random variable is a function that assigns a real value to each outcome of an experiment: X : S —> **R**

Example

How to define.

- Experiment: flipping a coin twice
- $S = \{\overset{2}{HH}, \overset{1}{HT}, \overset{1}{TH}, \overset{0}{TT}\}$
- Let X = number of heads
    - What are the assigned numerical values to the outcomes?

$$\frac{2+1+1+0}{2+2+2+2} = \frac{4}{8} = \frac{1}{2}$$

Types of random variables

离散 有限

- Discrete and finite nominal
  - Nominal — categorical with no implicit order (sample space: male, female)
  - Ordinal – ordered categories (sample space: small, medium, large)
- Discrete and countably infinite
  - Sample space: 1,2,3,… cases of coronavirus
- Continuous
  - Sample space: all possible values within a given interval
  - Theoretically at least. We don’t measure anything with infinite precision.
  - For example: systolic blood pressure

Notation

- X denotes the random variable
- x denotes a value of the random variable X
- Studies usually contain multiple individuals, and so a study generates observed values of random variables for each individual. When we are considering the possibility of multiple similar studies, a single study (with observed values for random variables for multiple individuals) is termed a "realization" of that study.
- As notation, if the random variable is named Y, when we are speaking about it in general we denote it by Y, whereas what we observe is Y=y. For example, Y=140.

Notation

- For our purposes, it is usually sufficient to discuss the values of a random variable for a single individual. However, if it is necessary to distinguish a value of 140 from participant 1 from a value of 150 from participant 2, we use the notation Y_1=140 and Y_2=150.

Events 事件.

- The result of an experiment is called an event

- Events are subsets of the sample space

- It's OK for an event to include the entire sample space, in which case it's certain to occur, or none of the elements of the sample space, in which case it's certain not to occur. Of course, the interesting cases fall between these two extremes.

- Examples of event after flipping a coin twice?

Probability

- The probability of an event E: $p = P(E) \in [0,1]$ is a real number representing our degree of certainty that E will occur
  - If $P(E) = 1$, then E is absolutely certain to occur
  - If $P(E) = 0$, then E is absolutely certain NOT to occur
  - If $P(E) = 0.5$, then the event E is equally likely to occur or not to occur
  - How do we interpret other values of p?

Classical interpretation of probability

- **Definition**: Event is a collection of outcomes, denoted by A,B,E etc. The probability that event E occurs is denoted by P(E). When all outcomes are equally likely, then $P(E) = \dfrac{\text{No. of outcomes}}{\text{No. of possible outcomes}}$
- Example: Flip a coin one time, what is the chance of getting a Head (H) ?
- Answer: Need to first ask whether the coin is fair? If the coin is fair, then P(H) = 1/2.

Relative frequency concept of probability (Empirical approach)

- If an experiment is repeated n times under essentially identical conditions, and if the event E occurs m times then P(E) ≈ m/n.
- E.g., flip the coin (since we don't know whether it is fair or not) a very large number of times and count number of H out of the total number of flips.

$$P(E) \approx \frac{\text{No. of outcomes in E}}{\text{No. of trials}}$$

- Example: if we flip the given coin 10,000 times and observe 4555 heads and 5445 tails, then for that coin, P(H) ≈ 0.4555.

Subjective probability

- Subjective probability reflects personal belief which involves personal judgment, information, intuition, etc.
- Example: What is the probability that we will get an effective COVID-19 vaccine this year?
- Each person may have different answers to the question.

Questions

HT HH TT TH

- Find the probability that exactly one head appears in two flips of a fair coin.

4

- Find the probability that two heads appear in two flips of a fair coin.
- Find the probability that the sum of two faces is greater than or equal to 10 when one rolls a pair of fair dice.

> = 1

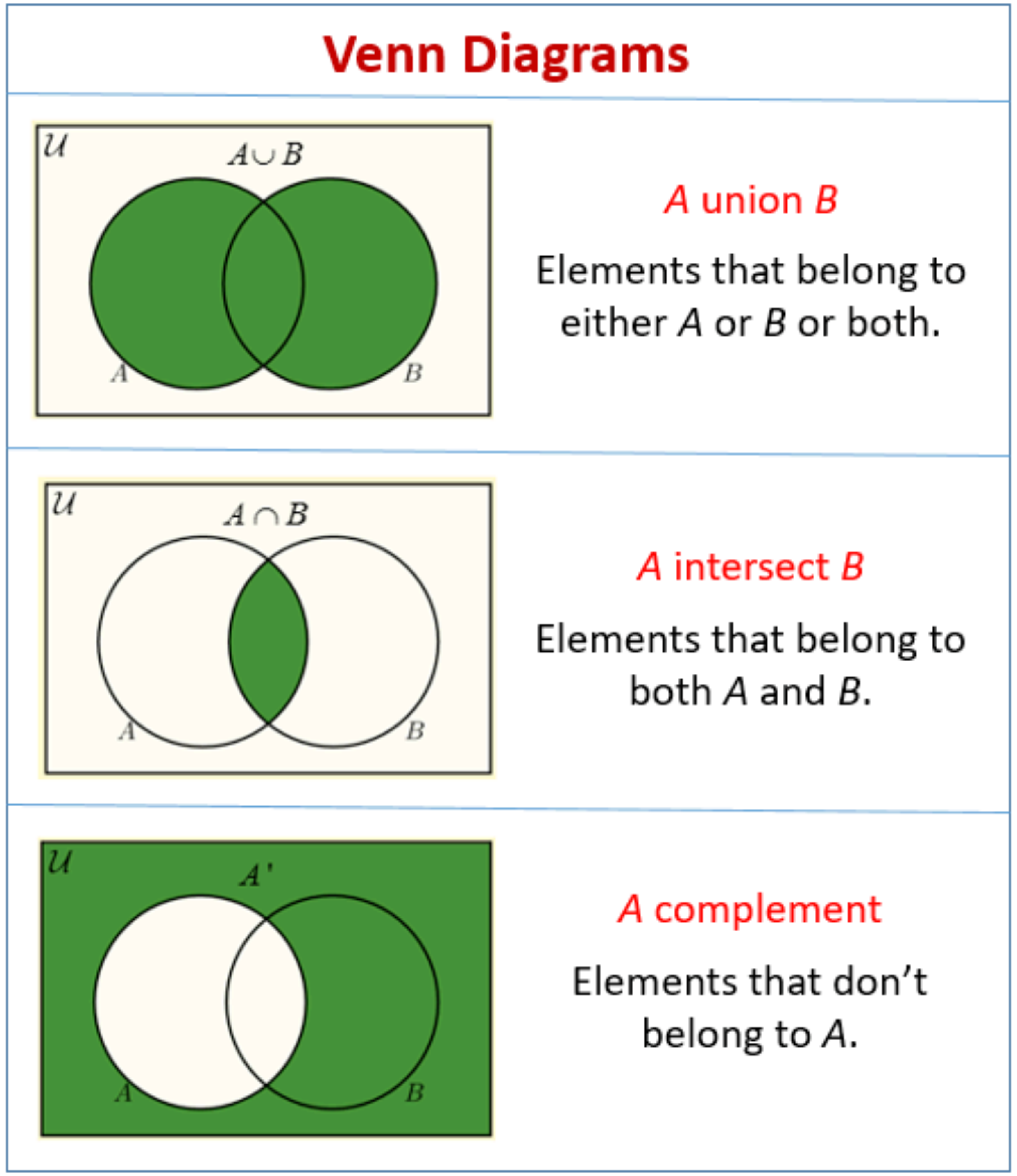
4 + 6 5 + 5

6 + 4 5 + 5

$$\frac{1+2+3}{36} = \frac{1}{6}$$

Probability laws (set operations)

- Union: $A \cup B$ = Outcomes in A or B or both
- Intersection: $A \cap B$ = Outcomes in both A and B
- Complement $\bar{A}$ or A^c = Outcomes not in A
- **Definition**: A and B are called mutually exclusive if the occurrence of outcomes in A excludes the occurrence of outcomes in B. In other words there are no elements in $A \cap B$ and thus $P(A \cap B) = 0$.
- A trivial example, A and A^c are mutually exclusive.



Venn Diagrams

A union *B*

Elements that belong to either *A* or *B* or both.

A intersect *B*

Elements that belong to both *A* and *B*.

A complement

Elements that don't belong to *A*.

Probability's properties

(a) $P(A^c) = 0.4$

(c) $= P(B) - P(A \cap B) = 0.3$

(d) $= P(A) + P(B) - P(A \cap B)$
$= 0.6 + 0.5 - 0.2$
$= 0.9$

(b) $P(A \cap B^c) = P(A) - P(A \cap B)$
$= 0.6 - 0.2 = 0.4$

- $0 \leq P(A) \leq 1$
- $P(A) = 1 - P(A^c)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Question: Given $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cap B) = 0.2$. Find (a) $P(A^c)$; (b) Find $P(A \cap B^c)$; (c) Find $P(B \cap A^c)$; (d) Find $P(A \cup B)$.

Example

$$\begin{array}{ccc} 1 & 1 & 2 \\ B & B & R \end{array}$$

- There are 3 balls in the urn where the first two are blue, and the third one is red, and the first two are numbered as 1 and the third one is numbered as 2. (a) One ball is drawn from the urn. Find the probability that the number on that ball is 1. (b) If someone tells you that the ball is blue, find the probability that the number on that ball is 1.

(a) $\frac{2}{3}$

(b) 1

Definitions

- $P(A \mid B) = P(A \cap B)/P(B)$
- $P(B \mid A) = P(A \cap B)/P(A)$
- $P(A \cap B) = P(B)P(A \mid B)$
- A and B are independent if and only if $P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$ or equivalently $P(A \cap B) = P(A) \cdot P(B)$. I.e., if A occurs, it has no effect on whether B occurs and vice versa.

Explanation of independence

- Independence has a substantive meaning and a mathematical meaning, with the latter being a representation of the former. The substantive meaning is that the events in question are physically and/or causally unrelated. If the science says "unrelated", the statistics can usually say "independent". Indeed, when independence is assumed rather than demonstrated it's because of the science.

- The mathematical definition is $P(A \,|\, B) = P(A)$, $P(B \,|\, A) = P(B)$ or equivalently $P(A \cap B) = P(A) \cdot P(B)$. If A and B are unrelated, knowing the value of B has no impact on the probability of A.

Explanation of independence

- Independence is sometimes assumed, and at other times it is assessed.
- As an example of assuming independence, consider 3 tosses of a fair coin, which is a form of sampling with replacement.
- The tosses are unrelated to one another, and so we can assume independence.
- The probability of 3 consecutive heads can be denoted as Pr{coin 1=H and coin 2=H and coin 3=H}, which, by independence, is simply the product of the probabilities of the 3 events: Pr{coin 1=H} * Pr{coin 2=H} * Pr{coin 3=H} = .5*.5*.5 = .125.

Conditional probability

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

- In general, $P(A\,|\,B) \neq P(A)$
- Current knowledge can change the sample space (possible outcomes)
- Example: A card is drawn from a well-shuffled deck. A = the card is King; B = the card is a face card (J,Q,K)
- P(A) = 4/52
- P(A|B) = 4/(3*4)
- Questions: (1) $P(B\,|\,A^c)$; (2) $P(B\,|\,A)$; (3) $P(A\,|\,B^c)$

Bayes' theorem

$P(B)$ conditioning on A.

- Knowing $P(B \mid A), P(B \mid A^c)$, and $P(A)$, is there a way to know $P(A \mid B)$?
- $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

$P(A \cap B) = P(A \mid B) \cdot P(B)$.

Bayes' theorem

- Knowing $P(B \mid A), P(B \mid A^c)$, and $P(A)$, is there a way to know $P(A \mid B)$?
- $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{P(A)P(B \mid A)}{P(B)}$

$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

Bayes' theorem

- Knowing $P(B|A), P(B|A^c),$ and $P(A)$, is there a way to know $P(A|B)$?
- $$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$
- $P(B) = P(A)P(B|A) + p(A^c)P(B|A^c)$

Medical testing

- Let D denote the event that an individual has the disease that we are testing for

- Let T+ denote the event that the test is positive, and T– denote the event that the test comes back negative

- $P(T+|D)$ is called the sensitivity of the test

- $P(T-|D^c)$ is called the specificity of the test

- Ideally, both $P(T+|D)$ and $P(T-|D^c)$ would equal 1. However, diagnostic tests are not perfect. They may give false positives and false negatives

Medical testing

$$P(T^+ \cap D) + P(T^+ \cap D^c) = P(T^+)$$

$$P(D \mid T^+) = \frac{P(D \cap T^+)}{P(T^+)}$$

- What is the probability that the tested person is infected if the test was positive?

- Give the test sensitivity is 0.98 and specificity is 0.995, and the disease prevalence is 1/300

$$= \frac{P(D) \cdot P(T^+ \mid D)}{\boxed{P(T^+)}}$$

$P(T^+ / D)$ sensitivity $= 0.98$

$P(T^- / D^c) = 0.995$

$$P(T^+ \mid D) \cdot P(D) + P(T^+ \mid D^c) \cdot P(D^c)$$

Independent and mutually exclusive

- If A and B are **mutually exclusive**, there is nothing in $A \cap B$, and $P(A \cap B)$ is 0. Thus, in general $P(A \cap B) = 0 \neq P(A) \cdot P(B)$
- Mutually exclusive events are dependent in general.

互斥

相关.

② $P(A)+P(B)-1 = 0.9+0.7-1 = 0.6$

$\min(P(A), P(B)) = 0.7$

Example

$P(A)$ $P(B)$

- Let A and B be the following two happy events. A: get a job, B: buy a new car. It is given that P(A) = 0.9, P(B) = 0.7. What is the probability of double happiness: that is you get a job and buy a new car?

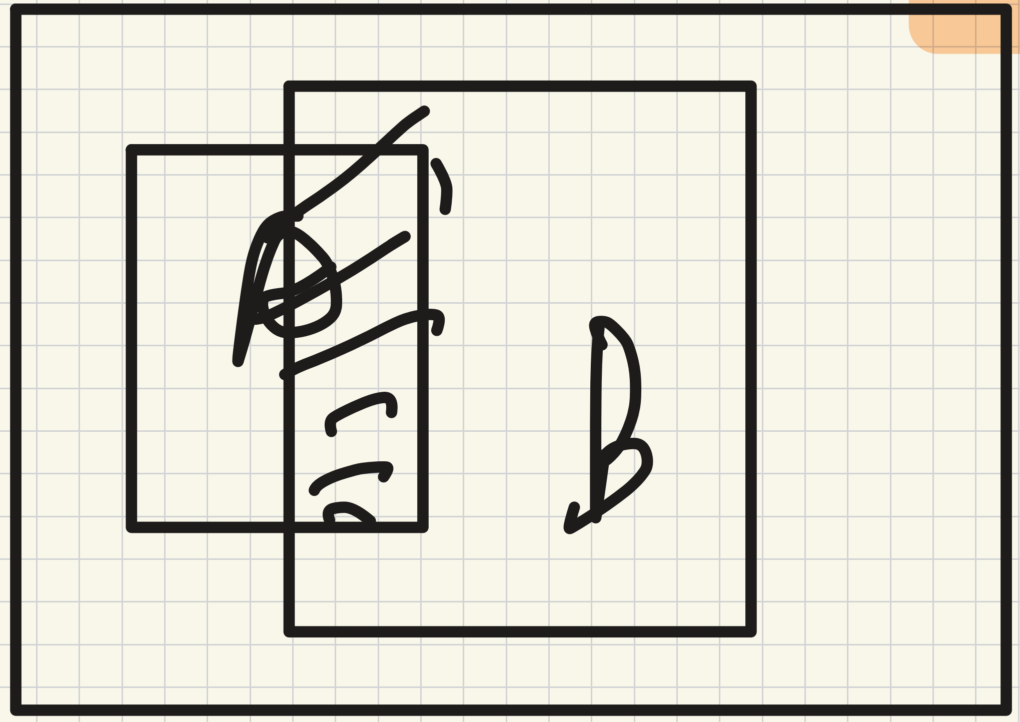
① independent: $P(A \cap B) = P(A) \cdot P(B)$

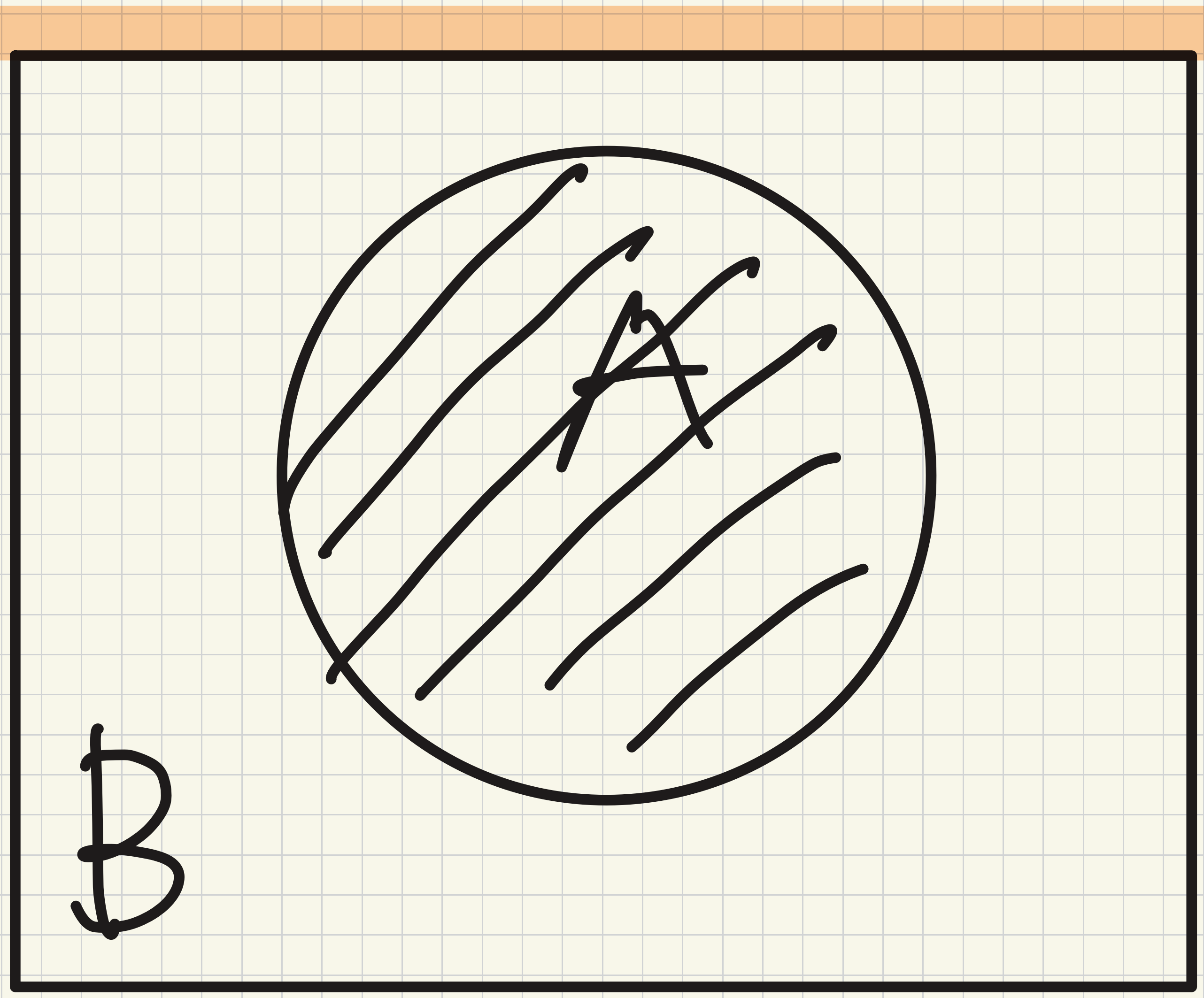
$= 0.9 \times 0.7$

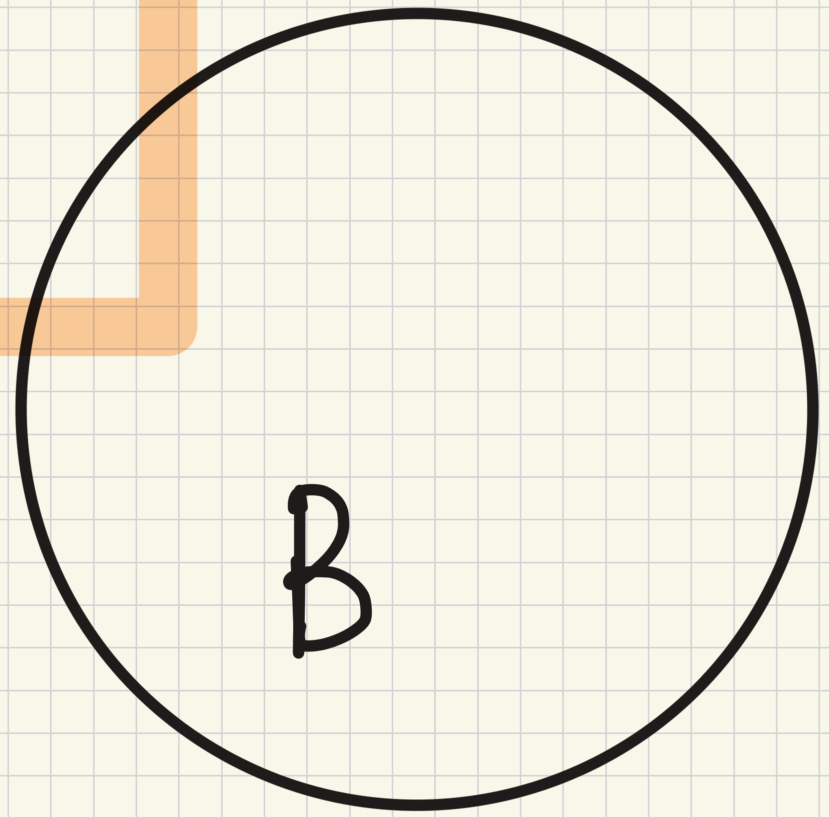
$= 0.63$

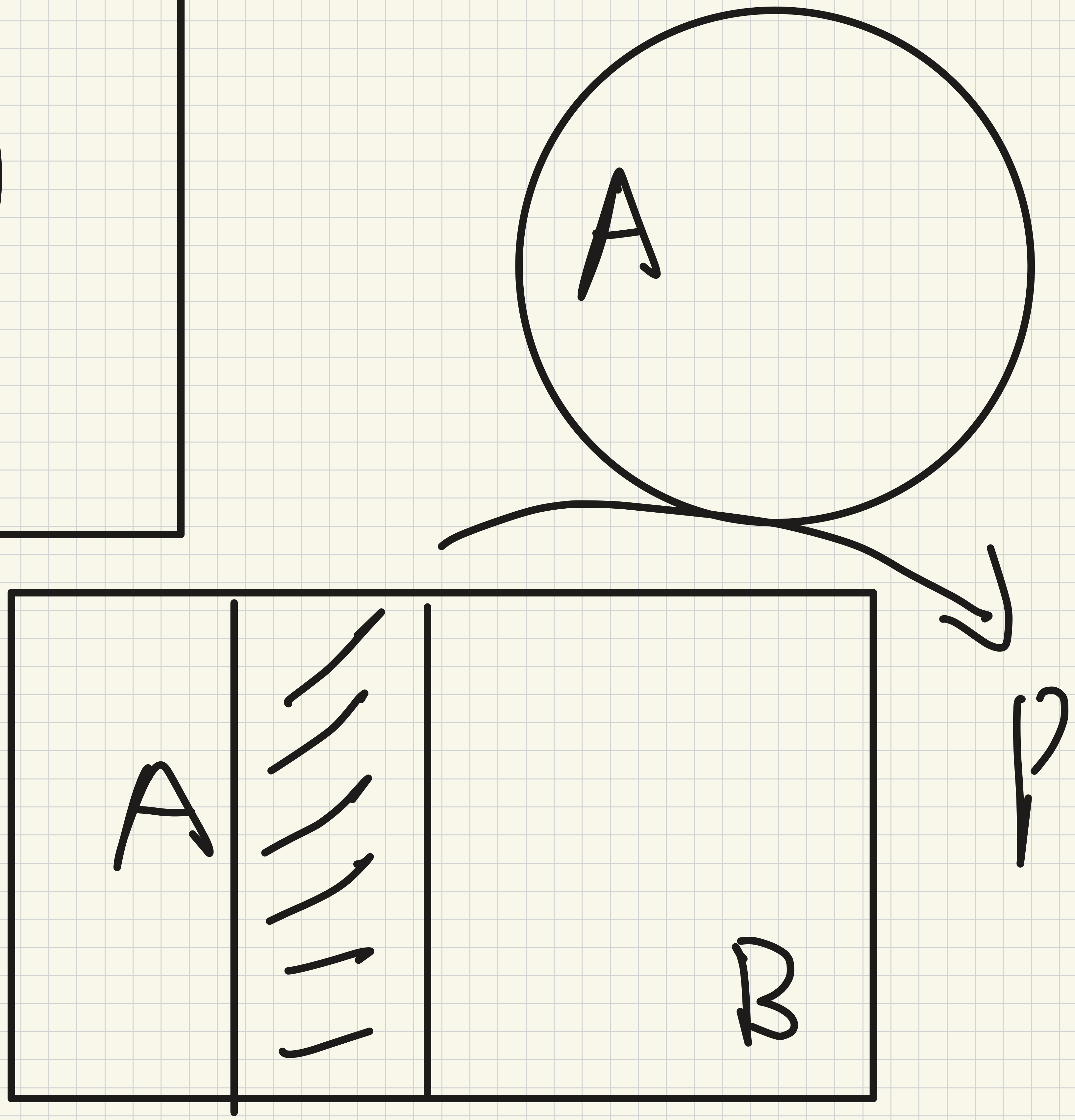
$$P(A\cup B) = P(A) + P(B) - P(A\cap B) \leq 1$$

$$\boxed{P(A\cap B) \geq P(A)+P(B)-1}$$

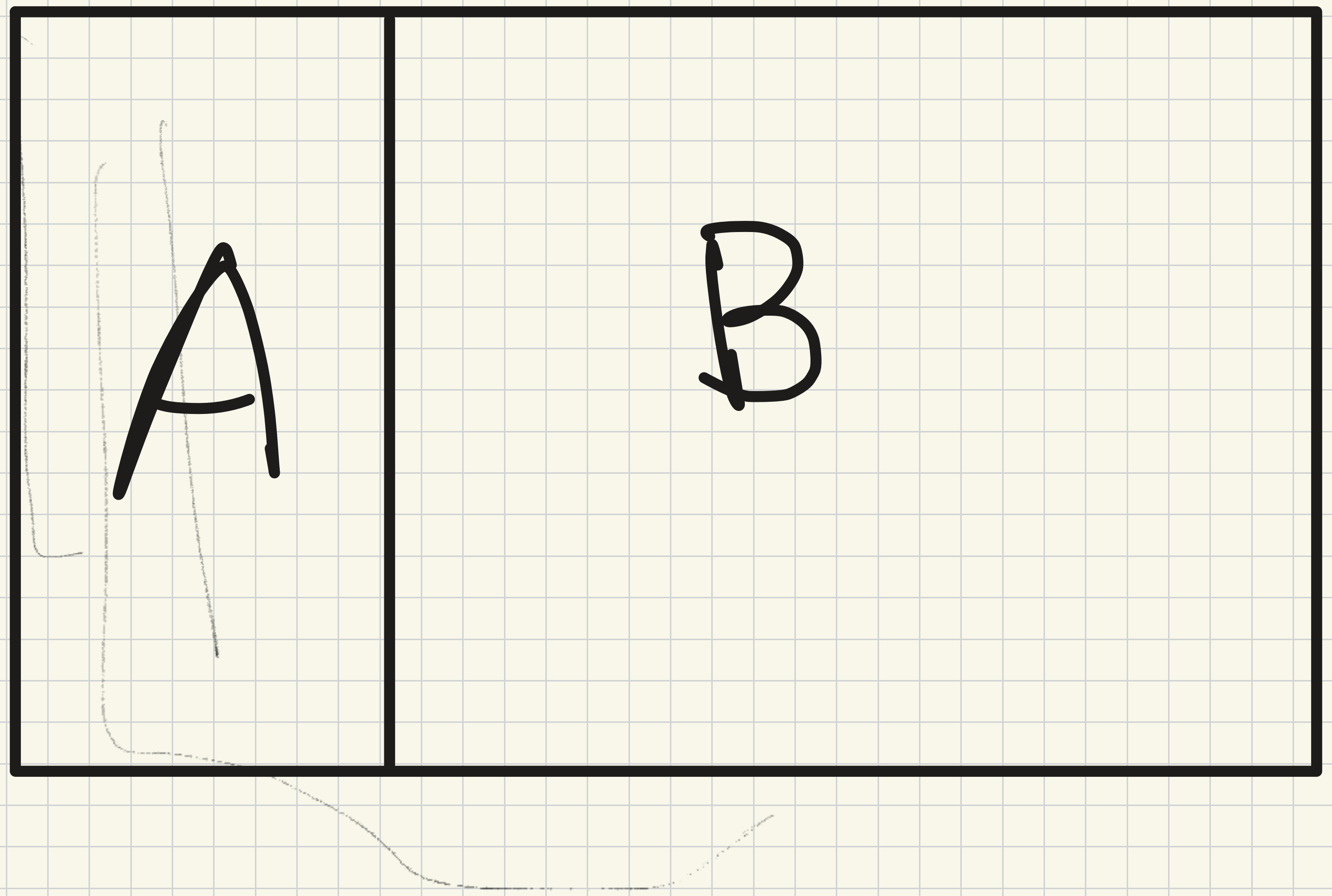








$P(A)+P(B)-1$



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

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

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

defination: $P(X \cap S) = P(X) \cdot P(S)$

Prove independence:

$P(X|S) \cdot \cancel{P(S)} = P(X) \cdot \cancel{P(S)}$

Example

independent: ② $P(X|S) = P(X)$

$P(X|S) = P(X)$

- Suppose that you throw a pair of fair dice. Denote by X the first outcome, and denote by Y the second outcome, and let S = X + Y. Is X independent of S? Is X independent of S = 7?

1+6

2+5

3+4

$6/36 = \frac{1}{6}$

$\frac{1}{6} = \frac{1}{6}$ ✓

① $P(S) \neq P(X|S)$

$X=6$ $S=5$ $P(S) \neq 0$.

Remark

推断

- Note that if the data are not the whole population but represent a random sample from a certain target population and we want to draw inference about whether the employment status and gender are related for the population, then we know that the sample may not follow the exact relationship $P(A \cap B) = P(A) \cdot P(B)$ even if such relationship holds for the population due to sampling variability.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- When we talk about the inference about the independence of a two way table, we will provide the rationale for the chi-square test for independence which measures how far off are the observations from being independent.

Basic principle of counting

- If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third outcomes, and if, ..., then there is a total of $n_1 n_2 \cdots n_r$ possible outcomes of the r experiments.

Arranging distinguishable objects

[illegible]

- Suppose that we have n distinguishable objects (e.g. the numbers 1,2,...,n). How many ways to order them (permutation) are there? If we have three objects a, b, c then the answer is 6: abc, acb, bac, bca, cab, cba.
- In general, there are n choices for the first object in our ordering. Then, we have n – 1 regardless of the first choice, and then we have n – 2 choices,... So there are in total $n(n-1)\cdots 2\cdot 1 = n!$ different orderings.

Arrangements when we only choose some

- Suppose that we have n distinguishable objects (e.g. the numbers 1,2,...,n). How many arrangements are there if we choose m (m < n) from them?

$$n \times (n-1) \times \dots \times (n-m)$$

$$= \frac{n!}{(n-m)!}$$

Arrangements when we only choose some

- Suppose that we have n distinguishable objects (e.g. the numbers 1,2,...,n). How many arrangements are there if we choose m (m < n) from them?
- Answer: $n \cdot (n-1) \cdots (n-m+1) = \dfrac{n!}{(n-m)!}$

Arrangements when we only choose some

- Suppose that we have n distinguishable objects (e.g. the numbers 1,2,...,n). How many arrangements are there if we choose m (m < n) from them?
- Answer: $n \cdot (n-1)\cdots(n-m+1) = \dfrac{n!}{(n-m)!}$
- **Another question**: Suppose we have n distinguishable objects. How many choices are there if we choose m (m < n) from them? This means that the order of the arrangement/choice does not matter.

Arrangements when we only choose some

- Suppose that we have n distinguishable objects (e.g. the numbers 1,2,...,n). How many arrangements are there if we choose m (m < n) from them?

P

- Answer: $n \cdot (n-1)\cdots(n-m+1) = \dfrac{n!}{(n-m)!}$

- **Another question**: Suppose we have n distinguishable objects. How many choices are there if we choose m (m < n) from them? This means that the order of the arrangement/choice does not matter.

C

- Answer: $\dbinom{n}{m} = \dfrac{n!}{m!(n-m)!}$