CLASS 1: THE ART OF COUNTING

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

Which kind of things are you expected to be able to count?

- **Permutations with repetition:** choosing r items in order (maybe repeated) from a set with n different elements can be done in n^r different ways when the order matters. For instance, in an alphabet of 26 words there are 26^5 words with exactly 5 letters. Observe that the order matters.

- **Permutations without repetition:** now, order matters and repetitions are not allowed; you must again choose (in order) r items from a set of n elements; in this case you get
$$\frac{n!}{(n-r)!} = n(n-1)\cdots(n-r+1).$$
For instance, in the case where you want to form a committee with 3 distinguishable positions (say president, vice-president and secretary) the number of ways this can be done is just
$$n(n-1)(n-2).$$

- **Combinations with repetition:** now, the order does not matter; you must select r items (maybe repeated) from a set with n different elements, but every item can be chosen more than once. This is equivalent to solve in non-negative integers the equation
$$x_1 + \ldots + x_n = r,$$
and you obtain
$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$
solutions.
For instance, you want to distribute $r = 3$ donuts between 14 students (not necessarily distinct). This can be done in $\binom{14+3-1}{3}$ ways.

- **Combinations without repetition:** order does not matter and repetitions are not allowed in this last case. Again, select r items from a set of n elements; this can be done in$\binom{n}{r}$. For example, if a teacher wants to pass exactly 4 students in a class of 14, this can be done in $\binom{14}{4} = \frac{14\cdot 13\cdot 12\cdot 11}{24}$ ways.

Inclusion-exclusion principle is a very powerful technique you must be able to use. For the case of 3 sets, if you want to determine the number of people that are in one (or more) of the three sets A, B and C (union), you use that
$$|A\cup B\cup C| = |A| + |B| + |C| - |A\cap B| - |A\cap C| - |B\cap C| + |A\cap B\cap C|.$$

Problem 1. Find the number of binary words of length 12 that begin with two zeros or finish with three zeros.

Problem 2. Find the number of non-negative integers less than 10^6 that do not contain the digit 2.

Problem 3. 12 couples are seated in a round table, and each people has his partner by his side. Determine of how many different ways people can sit in the following cases:

1. The seats are numbered.
2. The seats are not numbered.

Problem 4. A club has 30 members. Find the number of ways of doing the following selections:

1. Find 5 members to form the executive committee.
2. Choose a president, a vice-president, a secretary, a treasurer and a spokesman, if the positions are not compatible.
3. Choose a president, two vice-presidents, three secretaries, a treasurer and a spokesman, if the positions are not compatible.

Idea. For the first item, order does not matter (combinations) and repetitions are not allowed; you directly get $\binom{30}{5}$.
For the second, order does matter and repetitions are not allowed. You get $30 \cdot 29 \cdot 28 \cdot 27 \cdot 26$.
For the third one, you must take of repetitions in some places. The result is

$$30 \cdot \binom{29}{2} \cdot \binom{27}{3} \cdot 24 \cdot 23.$$

Problem 5. Determine the number of words with 8 letters that can be formed using an alphabet of 26 letters without repeating any letter and such that the words have at least 3 vowels.

Idea. You either use 5 vowels (and consequently 3 consonants you must choose), and then there are $\binom{21}{3} \cdot 8!$ options; or you use 4 vowels (you now also choose vowels) and 4 consonants, that can be done in $\binom{5}{4} \cdot \binom{21}{4} \cdot 8!$; or you use 3 vowels and 5 consonants, in total $\binom{5}{3} \cdot \binom{21}{5} \cdot 8!$ more ways.

Problem 6. Give proofs (better if they are combinatorial) of the following identities:

1. $\sum_{r=0}^{n-k} \binom{n-1-r}{k-1} = \binom{n}{k}$.
2. $k\binom{n}{k} = n\binom{n-1}{k-1}$ for $0 \leq k \leq n$.
3. $\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}$.
4. $\binom{\binom{n}{2}}{2} = 3\binom{n}{4} + 3\binom{n}{3}$ for $n \geq 4$.

Problem 7. Find the number of non-negative solutions of $x + y + z = 20$. Repeat the calculations if we impose that either $x \leq 10$ or $y \leq 3$.

Problem 8. 10 businessmen arrive to a very important meeting with their umbrella, that they leave in a box near the door. When they finish, they take a random umbrella from the box. What is the probability that everybody goes home with a different umbrella than the one he had at the beginning?

CLASS 2: AN INVITATION TO PROBABILITY

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

Which kind of things are you expected to be able to know?

- **Laplace law** (for simpler cases in which each elementary event has the same probability, like tossing fair coins and dices)**:** "the probability of an event A is the ratio of the number of cases favorable to it, to the number of all possible cases provided that the occurrence of these cases is equally possible".
- What is a probability space? What is an event in a discrete probability space? In general, each random experiment has associated a set E of all possible outcomes: this is the **sample space**. Their elements are the **elementary events** and its subsets are called **events**. If the cardinality of E is n, the number of events is 2^n. We can formulate this in a more abstract framework, talking about a pair (sometimes called an experiment) (E, A), where E is a non-empty set and A what is called a σ-algebra of sets of E; however, when E is finite both definitions agree. A **probability space** consists in assigning to each element X of A (alternatively here, to each subset of E) a real number between 0 and 1, say $\mu(X)$ in such a way that $\mu(E) = 1$ and $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $A \cap B = \emptyset$ (for infinite cases this definition must be adapted).
- **Independence:** two events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- **Conditioned probability:** The probability of A given B is
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
- **Law of composite probability:** Let $A_1, \dots, A_n$ be events with $P(\cap_{i=1}^{n-1} A_i) \neq 0$. Then,
$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n| \cap_{i=1}^{n-1} A_i).$$
- **Law of total probability:** Let $A, H_1, \dots, H_n$ be events such that $H_i \cap H_j = \emptyset$ for all $i, j \in [n]$ and such that $H_1 \cup \dots \cup H_n$ is the whole space. Then,
$$P(A) = \sum_{i=1}^{n} P(A \cap H_i) = \sum_{i=1}^{n} P(A|H_i)P(H_i).$$
- **Bayes' formula:** Under the above hypothesis,
$$P(H_k|A) = \frac{P(A|H_k)P(H_k)}{\sum_{i=1}^{n} P(A|H_i)P(H_i)}.$$

Let us work one of the prototypical examples with balls and urns. Imagine that we have 3 urns and different balls, that can be either blue (B) or red R. The distribution in boxes is: in box 1, 1 blue and 1 red; in box 2, 1 blue and 2 red; in box 3, 2 blue and 0 red.

Let us first calculate the probability of getting a blue ball. The probability of getting

a blue ball from the first box is $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$; the probability of getting a blue ball from the second box is $\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$; the probability of getting a blue ball from the third box is $\frac{1}{3} \cdot \frac{2}{2} = \frac{1}{3}$. Adding up, we have that the probability of getting a blue ball is $11/18$.
A natural question is: what is the probability of having chosen the first box provided that the resulting ball is blue? For that, we just use the result about conditioned probability, and we get

$$\frac{1/3 \cdot 1/2}{11/18} = \frac{3}{11}.$$

We can do the same with the second box and the result would be

$$\frac{1/3 \cdot 1/3}{11/18} = \frac{2}{11}.$$

Finally, for the third one, the probability is

$$\frac{1/3}{11/18} = \frac{6}{11}.$$

As you may see, the three probabilities add up one, $3/11 + 2/11 + 6/11 = 1$.

Problem 1. We roll two dices. Determine the probability of the following events:

1. We obtain a 6 in exactly one of the dices.
2. The two numbers are even.
3. The sum of the points is 4.
4. The sum is a multiple of 3.

Problem 2. If $\Pr(A) = 1/2$ and $\Pr(B) = 2/3$, give lower and upper bounds for $\Pr(A \cap B)$.

Problem 3. A total of n independent tosses of a coin that lands on heads with probability 0.038 is made. How large must n be so that the probability of obtaining at least one head is at least $1/2$.

Problem 4. Determine the numbers of children a couple must have if they want that at least is a woman with probability ≥ 0.95.

Problem 5. Consider a group of n different people.

1. Determine the probability that at least two have the same birthday.
2. Find n if we want that the previous probability is ≥ 0.5.

Problem 6. We have two urns. Urn number one has 2 white balls and 3 blue balls. Urn number 2 contains 3 white balls and 4 blue balls. We select one ball from urn 1 and put it in urn 2. Then, we select a ball from urn 2. Find the probability that it is blue.

Problem 7. We have 52 cards from a deck. Determine if the events

$$A = \{\text{the first card is a 7}\}$$

and

$$B = \{\text{the second card is a diamond}\}$$

are independent or not.

Problem 8. We divide the 52 cards from a deck into 4 players. Find the probability that each player has a 7.

Problem 9. What is the probability of getting a 6 when rolling a dice? What is this probability if I know that the result is even? And if I know that the result is greater or equal than 5?
A certain partner has 5 children. We know that the three younger are men. What is the probability that all are men?

Problem 10. We have a box with three coins of which two are faked. The probabilities of obtaining a head when flipping faked coins are 30% and 60%, respectively. A coin is chosen at random and then it is flipped three times, appearing two faces and one tail in this order. Find the probability that the flipped coin was not faked.

CLASS 3: AN INTRODUCTION TO RANDOM VARIABLES

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

Which kind of things are you expected to know?

- Let (E, A) be an experiment. A map $X : E \to \mathbb{R}$ is a (real) random variable (provided that it satisfies the "technical" condition that $X^{-1}((b, +\infty)) \in A$ for all real number b. Roughly speaking, to each possible result we associate it a real value. For instance, if we roll a dice we associate to each result the number we see in the face ($1, 2, 3, 4, 5$ or 6); if we flip a coin, we give the value 0 to the result "tail" and the value 1 to the result "head".
- Let X be a given discrete random variable. The probability function of X is the function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = P(X^{-1}(x)) = P(X = x)$. That is, we associate to each result a certain probability.
- "Naïve" examples: any constant function is a random variable. When E is finite, any (real-valued) function is a random variable. When we toss a coin, the random variable that assigns 1 to the (elementary) event "obtaining a head" and 0 to the even "obtaining a tail" is a random variable. When we roll a dice, the functions that assigns to each face its value is a random variable. In these cases, the probability functions are given in the case of the coin by the assignation of the value $1/2$ to both 0 and 1, and in the case of the dice by the assignation of the value $1/6$ to $1, 2, 3, 4, 5$ and 6.
- We now give more examples of random variables:
  - **Uniform:** $X \sim U(n)$ if $\mathrm{Im}(X) = \{x_1, \dots, x_n\}$ and $P[X = x_i] = 1/n$.
  - **Bernoulli:** $X \sim \mathrm{Ber}(p)$ if $\mathrm{Im}(X) = \{0, 1\}$, $P[X = 1] = p$ and $P[X = 0] = 1 - p$. When the variable takes the value 1 we talk about a success.
  - **Binomial:** $X \sim B(n, p)$ if $\mathrm{Im}(X) = \{0, 1, \dots, n\}$, $P[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$. It can be thought as the sum of n Bernoullis. For instance, when we throw n coins and we want to count the number of faces.
  - **Geometric:** $X \sim \mathrm{Geo}(p)$ if $\mathrm{Im}(X) = \{1, 2, \dots\}$, $P[X = k] = (1 - p)^{k-1} p$. It can be thought as the number of times a Bernoulli distribution must be repeated until we get a success.
- **Expectation:** the expectation of a discrete random variable X is $E(X) = \sum_{x \in \mathrm{Im}(X)} x P[X = x]$. We sometimes denoted by μ.

  With the previous notations, for a uniform random variable, $E(X) = (\sum_{i=1}^n x_i)/n$; for a Bernoulli, $E(X) = p$; for a binomial, $E(X) = np$; for a geometric, $E(X) = 1/p$.

  Expectation is linear! When X and Y are random variable (over the same probability space), and $a, b \in \mathbb{R}$,

  $$E(aX + bY) = aE(X) + bE(Y).$$
- **Variance:** $\mathrm{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$.

Problem 1. A player rolls a dice, gets a 6 and wins. What is the probability that he cheated, assuming that 40% of players are cheaters?

Problem 2. Suppose that the incidence of AIDS in the population of a certain country is 0.1%, and that the probability that a test correctly detects the presence of the virus is 95%. Conversely, the probability of a false positive is a 1%. A person that has done the AIDS test has been said that he is ill. Determine the probability that he is ill.

Problem 3. There are three urns with the following compositions:

$$U_1(3W, 4B), \quad U_2(2W, 3B), \quad U_3(2W, 4B).$$

Two dices are rolled. If the sum obtained is smaller or equal than 5, we choose the first urn. If the sum is 6 we choose U_2; otherwise we choose U_3. Then, a ball is picked up from the chosen urn. First of all, find the probability of being white. Then, knowing that it is white, find the probability that the sum of points obtained in the dices were 6.

Problem 4. Consider a random permutation of $[n]$. Determine the expected number of fixed points and the variance in the expected number of fixed points.

Problem 5. Let X be a random variable with probability function

$$\Pr(X = k) = \frac{1}{k(k+1)},$$

for $k = 1, 2, \ldots$. Prove that it defines a probability function and determine for which $\alpha \in \mathbb{R}$, $\mathbb{E}(X^\alpha) < \infty$.

Problem 6. We roll n dices. X is the random variable counting the number of pairs with the same result. Give the expectation and the variance of X.

Problem 7. A collector buys closed envelopes that contain one of the n cards of a collection with the same probability. Determine the expected number of envelopes he must buy to have the complete collection.

Problem 8. Determine the expected number of trials if we want to obtain n faces in a row when tossing a fair coin. Repeat the calculation if it is a faked coin with probability p of having a face.

CLASS 4: A REVIEW OF PROBABILITY

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

Problem 1. In a TV show, participants must choose at random one box from a total of six (numbered from 1 to 6). Inside the box, they will find a question that they must answer to win a prize. Boxes 1, 2 and 3 contain questions of "medium" difficulty (0.5 probability of success); boxes 4 and 5 contain "hard" questions (0.2 probability of success); and box 6 contains "easy" questions (0.8 probability of success).

- What is the probability of having a "hard" question?
- What is the probability of having a success?
- If we know that we have succeeded, what is the probability of having chosen one of the three first boxes?

Idea. We give the numerical solutions:

- $3/6 = 1/2$.
- $3/6 \cdot 1/2 + 2/6 \cdot 1/5 + 1/6 \cdot 4/5 = 1/4 + 1/15 + 2/15 = 9/20$.
- By conditioned probability, $\frac{1/4}{9/20} = \frac{5}{9}$.

Problem 2. A certain random variable X takes 3 different values, $1, 2, 3$, with the following probability function

$$\Pr(X = 1) = 3k/2, \quad \Pr(X = 2) = 1 - 2k, \quad \Pr(X = 3) = k/2,$$

where k is a certain real number.

- Find the range of values of k such that what we have defined is really a probability function.
- Find the expectation and the variance of X.
- What is the mode, that is, the value that X takes with higher probability?

Solution. Let us comment some important features about this problem:

- Two things must be satisfied: probabilities must add up one and each probability must be in the interval $[0, 1]$. For the sum being one, observe that

  $$\Pr(X = 1+) + \Pr(X = 2) + \Pr(X = 3) = 3k/2 + 1 - 2k + k/2 = 1,$$

  no matter which value k takes. The other conditions are

  $$0 \leq 3k/2 \leq 1 \Leftrightarrow k \geq 0 \text{ and } k \leq 2/3,$$

  $$0 \leq 1 - 2k \leq 1 \Leftrightarrow k \geq 0 \text{ and } k \leq 1/2,$$

  $$0 \leq k/2 \leq 1 \Leftrightarrow k \geq 0 \text{ and } k \leq 2.$$

  Then, we get that $k \geq 0$ and $k \leq \min\{2/3, 1/2, 1\} = 1/2$. We conclude that $k \in [0, 1/2]$.

- For the expectation,

  $$E(X) = 1 \cdot \Pr(X=1) + 2 \cdot \Pr(X=2) + 3 \cdot \Pr(X=3) = 3k/2 + 2 - 4k + 3k/2 = 2 - k.$$

  For the variance, we must compute the following quantities:

  $$(1 - E(X))^2 = (1 - (2-k))^2 = (k-1)^2 = k^2 - 2k + 1,$$

  $$(2 - E(X))^2 = (2 - (2-k))^2 = k^2,$$

  $$(3 - E(X))^2 = (3 - (2-k))^2 = (k+1)^2 = k^2 + 2k + 1.$$

  Now, by definition

  $$\mathrm{Var}(X) = \Pr(X=1) \cdot (1 - E(X))^2) + \Pr(X=2) \cdot (2 - E(X))^2 + \Pr(X=3) \cdot (3 - E(X))^2$$

  $$= 3k/2(k^2 - 2k + 1) + (1 - 2k)k^2 + k/2(k^2 + 2k + 1) = -k^2 + 2k.$$

- $\Pr(X=1)$ is always greater than $\Pr(X=3)$, so the most frequent value can only be either 1 or 2. In particular

  $$\Pr(X=1) > \Pr(X=2) \Leftrightarrow 3k/2 > 1 - 2k \Leftrightarrow k > 2/7,$$

  $$\Pr(X=2) > \Pr(X=1) \Leftrightarrow 3k/2 < 1 - 2k \Leftrightarrow k < 2/7.$$

  Hence, when $k \in [0, 2/7)$ the mode is 2; when $k \in (2/7, 1/2]$ the mode is 1; when $k = 2/7$ the values 1 and 2 have the same probability.

Problem 3. Determine the expected numbers of times you get a 6 when rolling a dice 100 times.

Solution. Let X_1 be the random variable that takes the value 1 if we have obtained a 6 in dice number 1 and 0 elsewhere. We clearly have $E(X_1) = 1/6$. Define in the same way $X_2, X_3, \ldots, X_{100}$. Then, we observe that the expected number of times you obtain 6 is

$$E(X_1 + X_2 + \ldots + X_{100}) = E(X_1) + E(X_2) + \ldots + E(X_{100}) = 100 \cdot 1/6 = 100/6 \simeq 16.7.$$

Problem 4. Determine the expectation in the number of times you through a (fair) coin until you get a head.

Solution. The number of times you must flip a coin that has probability p of being head (and then probability $1-p$ of being tail) is $1/p$. Hence, if $p = 1/2$ we obtain that it must be flipped twice.

Problem 5. Two players roll a dice alternatively until one get a 6. Determine the probability that the player who starts wins the game.

Idea. We get that the result is

$$\frac{1}{6} + \frac{1}{6}\Big(\frac{5}{6}\Big)^2 + \frac{1}{6}\Big(\frac{5}{6}\Big)^4 + \ldots = \frac{1/6}{1 - (5/6)^2} = \frac{6}{11}.$$

CLASS 6: AN INVITATION TO STATISTICS

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

The following words are frequently used in describing statistical data: **population**, **sample**, **variable**, **parameter**, ... There are three kinds of statistical variables:

- Qualitative.
- Discrete quantitative.
- Continuous quantitative.

Given a set of data, say $x_1, \dots, x_n$ we have the following parameters:

- **Arithmetic mean (or average):** $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$.
- **Geometric mean:** $\mathrm{GM} = \sqrt[n]{x_1 \cdots x_n}$.
- **Harmonic mean:** $\mathrm{HM} = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}$.
- **Quadratic mean:** $\mathrm{QM} = \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}$.
- **Trimmed arithmetic mean:** moderates the influence of the extreme values when computing the mean. The trimming percentage is the percentage of valued deleted from each end of the ordered list. The $\alpha(\%)$-trimmed mean is the arithmetic mean of data remaining in the sample after removing $\alpha/2(\%)$ of both the larges and smallest scores.
- **Median:** If n is odd, we just take the central value. When n is even we take the mean of the two central values.
- **Quartiles (Q_1 and Q_3):** the first quartile Q_1 is the value that is above 25% of the population and below the other 75%; the third quartile Q_1 is the value that is above 75% of the population and below the other 25%. When n is odd, we just split the data in two halves, each of them containing the central value, and then Q_1 and Q_3 are the median of each half. When n is even, we split the data in two again, but the two values that are taken into account for the median are distributed, one for each half (and then Q_1 and Q_3 are the median of each part).
- **Percentiles:** the percentile p_i is the value that leaves below it the $i\%$ of the population, and above it, the $(100 - i)\%$. Although there is no universal convention, typically, in an ordered set, for the p_i percentile in a set of n, we shall consider the value that occupies the first position greater or equal than $n \cdot \frac{i}{100}$.
- **Mode:** most repeated value.
- **Variance**: $\sigma^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$. There is a concept of quasi-variance or sample variance in which we divide by $n-1$ and not by n.
- **Standard deviation**: σ; square root of the variance.
- **Variation coefficient:** $\frac{\sigma}{\bar{x}}$.

- **Mean deviation:** $\frac{1}{n}\sum_{k=1}^{n}|x_k-\bar{x}|$.
- **Range:** $x_n - x_1$.
- **Interquantile range:** $Q_3 - Q_1$.

Data is frequently sorted in frequency tables. For that, we select the extreme values, say a and b, and determine the difference $r = b - a$. We decide the number of intervals we want to form, typically between 6 and 15. We form the intervals in such a way that the lower extreme of the first one is slightly smaller than a and the upper extreme of the last one is slightly greater than b. It is desirable that extremes of intervals do not agree with any of the data. We use the following terminology: **class interval**, **class limits**, **class size** and **class mark**.

When we have a distribution given by a frequency table, we can compute parameters such as the mean or the variance:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}; \qquad \sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2.$$

The **absolute frequency** of data x_i is the number of times n_i that x_i appears, and is denoted as $f_a(x_i) = n_i$; the **relative frequency** refers to the proportion. When the values can be sorted, we talk about **cumulative frequencies** (both absolute and relative).

In these cases, for computing the median, the quartiles and the percentiles, several formulas can be used, although the easiest way is to use the polygons of cumulative frequencies.

For the graphical presentation of the data, the following are used:

- **Bar chart:** for categorical or quantitative discrete data.
- **Pie chart:** as the bar chart, but they are more effective for summarizing data sets when there are not too many different categories.
- **Histogram:** either for discrete or continuous numerical data.
- **Frequency polygons:** helpful for comparing sets of data.
- **Pictograms:** effective for giving general impressions.

CLASS 7: LINEAR REGRESSION
MSL 109: BASICS OF STATISTICS
Óscar Rivero Salgado

A **bivariate data set** is a collection of pairs of values that results of the jointly observation of two measurable characteristics X and Y of a population. It has the form

$$M_{x,y} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}.$$

We can also talk about the concepts of **absolute frequency** and **relative frequency**, as in the univariate case.
Here, if $x_1, x_2, \dots, x_h$ are the h distinct values of the variable X and $y_1, y_2, \dots, y_k$ are the k distinct values of the variable Y in the sample, the **jointly distribution** of frequencies is the **crosstab** containing the frequencies (either absolute or relative) of the pairs (x_i, y_j).

Given a set of data (x_i, y_i) we define as the **covariance** the parameter

$$\sigma_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}.$$

The **correlation** between the two variables in a bidimensional distribution is given by

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where σ_{xy} is the covariance and σ_x, σ_y are the standard deviations of each variable. This *adimensional* variable is always between -1 and 1. If the correlation is perfect (points being in a line), then either $r = 1$ or $= -1$. If the correlation is strong, $|r|$ is close to 1; elsewhere, it is close to 0.

One of the usual objectives is to draw line that fit well our set of data. To determine the line that best follows the set (x_i, y_i), we do a computation using either linear algebra or multivariable calculus. In any case, we obtain that the slope of the line is

$$m_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$$

and in general the equation of this line (the one that minimizes the squares of the distances of the points to it) is

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}).$$

The slope is sometimes called **regression coefficient**.
This regression line is useful for making predictions. For instance, $\hat{y}(x_0)$ is the estimated value of y corresponding to $x = x_0$; $\hat{x}(y_0)$ is the estimated value of x given $y = y_0$. The greater $|r|$ the best the approximation is.
Observe that another possibility is to consider the regression line in which x is given in terms of y (regression line of X over Y), in which

$$x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}).$$

We can also give this line as

$$y = \bar{y} + \frac{\sigma_y^2}{\sigma_{xy}}(x - \bar{x}).$$

Observe that this equation is not the same as the other one: when the correlation is weak, the two lines form a certain angle (close to 90^{o} if r is close to 0); however, for a strong correlation ($|r|$ close to 1) the two lines are almost the same.

CLASS 9: CONTINUOUS RANDOM VARIABLES: THE NORMAL DISTRIBUTION

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

The **probability density function** of a continuous random variable X is a function $f : \mathbb{R} \to \mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,

$$P[a < X \leq b] = \int_a^b f(x)\, dx$$

and it satisfies the following two properties:

- $f(x) \geq 0$ for all $x \in \mathbb{R}$.
- The area under the curve is 1, that is, $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Roughly speaking, the probability of being between a and b is the area under the curve. On the other hand, we can consider the **cumulative distribution function**, that we usually denote by capital letters, say F. $F(x)$ is the probability of being below x:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(x)\, dx.$$

One of the most classical examples is $U[a, b]$, the **uniform distribution** that takes with the same probability any value between a and b. Both the expectation and the median are $\frac{a+b}{2}$. The variance is $\frac{(b-a)^2}{12}$.
Another example is the **exponential distribution** of parameter λ ($\lambda > 0$). It is given by $f(x) = \lambda e^{-\lambda x}$ when $x \geq 0$ (and 0 elsewhere). The expectation is $\frac{1}{\lambda}$, the variance is $\frac{1}{\lambda^2}$ and the median is $\frac{\log 2}{\lambda}$.
However, our more relevant example will be given by the **normal distribution** (also called gaussian). It is described with two parameters, the mean μ and the variance σ^2. The density function is given by

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

It is a symmetric distribution. If $X \sim N(\mu, \sigma)$ and we denote by

$$Z = \frac{X - \mu}{\sigma},$$

then $Z \sim N(0, 1)$.
The importance of the normal distribution comes from the **central limit theorem**: let $X_1, \ldots, X_n$ be n identical and independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$. Then,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

approaches to a normal distribution (this means that for a big n, we can be arbitrarily close to the normal).
In particular, we can use this for a binomial random variable, $X \sim B(n, p)$ and in this case

$$Z_n = \frac{X - np}{\sqrt{np(1-p)}}$$

will behave as a normal distribution for n big enough. Then, $X \sim N(np, \sqrt{np(1-p)})$.

CLASS 11: INTERVAL ESTIMATION

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

A **population** is a set of elements in which a characteristic or given variable is studied. Any subset of representative elements of a population is called a **sample**. To obtain a data from a population we can proceed in two ways, through a **census** (the whole population is studied) or by selecting a sample; this procedure is called **sampling**. There are several types of sampling:

1. Simple random sampling.
2. Systematic random sampling.
3. Stratified random sampling.
4. Random sampling by clusters.
5. Directed random sampling.
6. Non-random sampling by quotas.
7. Deliberate non-random sampling.

Given a population characterized by a random variable X, a random sample is any set $X_1, \dots, X_n$ of independent random variables, identically distributed, all with the same distribution and the same parameters as X. An **estimator** is a function of the sample appropriate to estimate a population parameters.
The **method of moments** consist in equating the sample moments with the population moments. Let us consider the example of a normal distribution, with known σ and unknown μ. Since $\mathbb{E}(X) = \mu$, we can estimate with $\hat{\mu} = \frac{\sum x_i}{n}$.
For a variable that is uniformly distributed in $[a, b]$ (a, b unknowns),

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \operatorname{Var}(X) = \frac{(b-a)^2}{12}.$$

Hence,

$$\hat{a} = \mathbb{E}(X) - \sqrt{3\operatorname{Var}(X)}, \qquad \hat{b} = \mathbb{E}(X) + \sqrt{3\operatorname{Var}(X)}.$$

It turns out that when the variance is unknown, the best way to estimate it is not the population variance we have studied, but the sample variance, namely

$$s^2 = \frac{n-1}{n}\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}.$$

When a random variable X follows a distribution of mean μ, we call **confidence interval** (or characteristic interval in some places) corresponding to a probability p to an interval centered at the mean, $(\mu - k, \mu + k)$ such that

$$P[\mu - k < X < \mu + k] = p.$$

We will study now the most prototypical situations concerned with normal populations. Let us begin with the **confidence interval for the mean with known variance**.

Let $X \sim N(\mu, \sigma^2)$, where μ is unknown and we want to obtain a confidence interval for μ with confidence level $1-\alpha$. That is, we want

$$P[l_1 \leq \bar{X} \leq l_2] = 1 - \alpha.$$

We know that the sample mean $\bar{X}$ is distributed as a normal random variable with parameters μ and σ^2/n. We can use the normal distribution and find values of l_1 and l_2 such that

$$P\Big(l_1 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq l_2\Big) = 1 - \alpha,$$

and hence we get that a confidence interval for the mean of a normal distribution with known variance is

$$\Big(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big).$$

Observe that $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is the maximum admissible error. It is directly proportional to σ and inversely proportional to the size of the sample. When the **variance is unknown**, a confidence interval for the population with $1-\alpha$ level of confidence is given by

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}.$$

Here, $t_{\alpha/2,n-1}$ is a t-Student distribution with $n-1$ degrees of freedom. When n is big (say greater than 30 we can approximate $t_{\alpha/2,n-1}$ by $z_{\alpha/2}$.

Let us continue by explaining the **estimation of the differences between the means of normal populations**. If $\bar{x}_1$ and $\bar{x}_2$ are the values of the means of two independent random samples of size n_1 and n_2 of two normal populations with known variances σ_1^2 and σ_2^2, then a confidence interval of confidence level $1-\alpha$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}}.$$

When the samples are large, we can use this same expression changing σ_i by s_i. However, the correct expression when the variance is unknown is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

If $X \sim \text{Bin}(n,p)$ and n is large, we can estimate $\hat{p} = \frac{x}{n}$. Then, a confidence interval for the population proportion is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

CLASS 12: HYPOTHESIS TESTING

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

A parameter can be estimated from sample data either by a single number, a point estimate or an entire interval of plausible values, a confidence interval. However, the objective of an investigation is usually to decide which of two contradictory claims about the parameter is correct. Methods for accomplishing this constitute the so called **hypothesis testing**. Let (X, f_X) be a distribution of probability and let $X_1, \dots, X_n$ be a random sample. A hypothesis test of parametric test about the population parameters consists of the following stages:

1. **The null hypothesis**, also called H_0, is the claim that is initially assumed to be true.

2. **An alternative hypothesis**, also denoted by H_1, is the assertion that is contradictory to the null hypothesis.

3. **A test statistic** is a function of the sample data on which the decision of reject H_0 or do not reject H_0 is to be based on.

4. **The significance level** of the test, denoted by α, and that is the probability that H_0 is rejected being true.

5. **A rejection region** is the set of all test statistic values for which H_0 will be rejected.

There are two types of errors: a **Type I error** occurs when H_0 is rejected being true; a **Type II error** occurs when H_0 is not rejected being false. The probability of a Type I error is denoted by α and the probability of a Type II error by β.

Let us see some examples about the population mean.
Upper-tailed test: here, the null hypothesis is $\mu = \mu_0$ and the alternative hypothesis is $\mu > \mu_0$. The statistical test is $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ and H_0 is rejected if $z > z_\alpha$. The **lower-tailed test** is done in an analogous way. The **two-tailed test** has as alternative hypothesis $\mu \neq \mu_0$ and hence H_0 is rejected when $|z| > z_{\alpha/2}$.

We know move to the tests for the **difference between means of two populations**. When we consider two independent random samples of size n_1 and n_2 of normal populations with means μ_1 and μ_2 and known variances σ_1^2 and σ_2^2, we may want to decide if $\mu_1 = \mu_2$. Here, the statistical test is

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

If the samples are small and the variances unknown we may use the t-Student distribution.
Let us finish by explaining the chi-squared tests. The aim in this type of tests is to compare proportions of two or more populations. The tests typically carried out are those of **goodness of fit** and **independence**. Suppose we perform an experiment such that their results can be classified into k categories of cells and that it is repeated

n times. Furthermore, assume that the odds or proportions of the different results are $p_1, \ldots, p_k$, with $\sum p_i = 1$ and that in the total of the n repetitions the frequencies observed were $O_1, \ldots, O_n$ with $\sum O_i = n$.
Then, the statistical test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i},$$

where $e_i = np_i$ is the expected frequency. We reject H_0 (the hypothesis that $\pi_1 = p_{10}, \ldots, \pi_k = p_{k0}$) when $\chi^2 > \chi^2_{\alpha,k-1}$.

CLASS 14: COURSE SUMMARY

MSL 109: BASICS OF STATISTICS

Óscar Rivero Salgado

Problem 1. We take two cards from a deck. If both are diamond, we automatically win a prize; if exactly one is diamond and the other not, we get a prize with 50% of probability. In case of obtaining no diamond, we will have the prize with a probability of 10%.

- What is the probability of obtaining exactly one diamond?
- What is the probability of winning the prize?
- If we know that we have won the prize, what is the probability of having obtained two diamonds?

Problem 2. A total of n independent tosses of a coin that lands on heads with probability 0.038 is made. How large must n be so that the probability of obtaining at least one head is at least 1/2.

Problem 3. The following sample corresponds to the bivariate variable (X, Y).

x **variable**	y **variable**
1.5	23.0
1.5	24.5
2.0	25.0
2.5	30.0
2.5	33.5
3.0	40.0
3.5	40.5
3.5	47.0
4.0	49.0

Find the correlation coefficient.

Determine the equation of the fitting line.

Predict the value of y when $x = 5.0$ and the value of x when $y = 70.0$.

Problem 4. The following table shows the frequency distribution of grades in a class of 100 students:

Height	**Frequency**
1	6
2	14
3	42
4	34
5	4

Determine the standard deviation and the quartiles (Q_1 and Q_3) of the distribution.

Problem 5. A teacher wants to estimate the mean weight of all his students with an error smaller than 1 cm using a sample of 400 children. Knowing that $\sigma = 6.0$ kg, determine the confidence level.
In a similar experiment, the teacher used a sample of 20 children and he estimated the value of the mean (70 kg) as well as the standard deviation (5 kg). Find a confidence interval given for a value of $\alpha = 0.05$.

Problem 6. A certain candidate (say A) affirms that he will get more than one third of the supports in the presidential election. We take a sample of 1000 people and 315 say that they will vote for A. Can we accept the candidate's hypothesis with a confidence level of 90%?