Lecture Notes for Math 372: Elementary Probability and Statistics

Last updated: January 19, 2026

Contents

0 Tentative course outline

This course is a problem-oriented introduction to the basic concepts of probability and statistics, providing a foundation for applications and further study.

1. **Weeks 1-2.** Sampling distributions (4 lessons).

   chi-squared, t, and F distributions, distributions of sample mean and variance

2. **Weeks 3-4.** Point estimation (5 lessons)

   properties and methods of point estimation

3. **Weeks 5-6.** Interval estimation (4 lessons)

   Confidence intervals for means, variances, proportions and differences

4. **Weeks 7-12.** Hypothesis Testing (19 lessons)

   Neyman-Pearson lemma, likelihood ratio test; tests concerning means and variances, tests based on count data, nonparametric tests, analysis of variance

5. **Weeks 13-14.** Regression and correlation (6 lessons)

   regression, bivariate normal distributions, method of least squares

1 2025-01-12 | Week 01 | Lecture 01

- give syllabus
- do activity with why you're in this course

> ***The nexus question of this lecture:*** *What is a probability?*

Reading assignment: Sections 1.1, 1.2, 1.3, 2.1, 2.4 of the textbook.

1.1 What is probability?

1.1.1 A general framework: sample space, events, etc

We begin with a general framework and some terminology to formalize the notions of probability. This is based on section 2.4 in the textbook.

- An ***experiment*** is an activity or process whose outcome is subject to uncertainty, and about which an observation is made.

  Examples include flipping a coin, rolling a dice, measuring the size of a wave, or the amount of rainfall, conducting a poll, performing a diagnostic test, opening a pack of Pokemon cards, etc.

- The ***sample space*** S of an experiment is the set of all possible outcomes. The elements of the sample space are called ***sample points***.

  We think of each sample point as representing a unique outcome of the experiment. In the case of rolling a dice, the sample points are $1, 2, 3, 4, 5$ and 6, and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

- We use the term ***event*** to refer to a collection of outcomes, i.e., a subset of S.

  Example: if our experiment is rolling a 6-sided dice, here are some events

$$\begin{aligned} A &= [\text{observe an odd number}] & E_2 &= [\text{observe a 2}] \\ B &= [\text{observe an even number}] & E_3 &= [\text{observe a 3}] \\ C &= [\text{observe a number less than 5}] & E_4 &= [\text{observe a 4}] \\ D &= [\text{observe a 2 or a 3}] & E_5 &= [\text{observe a 5}] \\ E_1 &= [\text{observe a 1}] & E_6 &= [\text{observe a 6}] \end{aligned}$$

- There are two types of events: ***compound events***, which can be decomposed into other events, and ***simple events***, which cannot.

  In the above example, the events A, B, C and D are compound events. $E_1, \ldots, E_6$ are simple events.

- A sample space is ***discrete*** if it is countable (i.e., finite or countably infinite). In a discrete sample space S, the set of all possible events is the *power set* of S.[1]

  In the dice-rolling example, the set of all possible events is $\{E : E \subseteq \{1, 2, 3, 4, 5, 6\}\}$.

$$\begin{aligned} A &= [\text{observe an odd number}] = \{1, 3, 5\} & E_2 &= [\text{observe a 2}] = \{2\} \\ B &= [\text{observe an even number}] = \{2, 4, 6\} & E_3 &= [\text{observe a 3}] = \{3\} \\ C &= [\text{observe a number less than 5}] = \{1, 2, 3, 4\} & E_4 &= [\text{observe a 4}] = \{4\} \\ D &= [\text{observe a 2 or a 3}] = \{2, 3\} & E_5 &= [\text{observe a 5}] = \{5\} \\ E_1 &= [\text{observe a 1}] = \{1\} & E_6 &= [\text{observe a 6}] = \{6\} \end{aligned}$$

[1] If S is not discrete, a complication arises: in that case, some subsets of S are too wild and untameable for us to treat them mathematically as "events". Resolving that issue requires introducing measure theory, which is beyond the scope of this class, so we will ignore it and simply steer clear of any setting where any issues might arise.

- Some observations about events:
  - The sample points are *elements* of S. The simple events are *singleton subsets* of S. In the dice example, we have:
    - Sample points: 1,2,3,4,5,6.
    - Simple events: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$.
  - The empty set $\emptyset$ and the whole sample space S are always both events: $\emptyset$ is the event "nothing happens" and S is the event "something happens".
  - Events satisfy the properties of a boolean algebra:
    - **"And":** If E and F are events, then $E \cap F$ is the event that E and F occur.
    - **"Or":** If E and F are events, then $E \cup F$ is the event that E or F occurs.
    - **"Not":** If E is an event, then $E^c = S \backslash E$ is the event that E does not occur.
- Two events E and F are ***mutually exclusive*** if $E \cap F = \emptyset$. This means that E and F cannot both happen at the same time.

  In the dice example, the events A and B are mutually exclusive, since the dice roll cannot be both even and odd. But A and C are not mutually exclusive because $A \cap C = \{1, 3\} \neq \emptyset$. If a 1 or a 3 is rolled, then both A and C occur.

1.1.2 Definition of probability measure

Definition 1 (Probability measure)**.** Let S be a sample space associated with an experiment. A function $\mathbb{P}$ is said to be a ***probability measure*** on S if it satisfies the following three axioms:

A.1 (Nonnegativity) For every event $E \subseteq S$,
$$\mathbb{P}[E] \geq 0.$$

A.2 (Total mass one) $\mathbb{P}[S] = 1$.

A.3 (Countable additivity) If $E_1, E_2, \ldots$ is a sequence of events which are pairwise mutually exclusive (meaning $E_i \cap E_j = \emptyset$ if $i \neq j$), then
$$\mathbb{P}[E_1 \cup E_2 \cup \ldots] = \sum_{i=1}^{\infty} \mathbb{P}[E_i].$$

If $\mathbb{P}$ is a probability measure, then for every event $E \subseteq S$, the number $\mathbb{P}[E]$ is called the ***probability*** of E.

The above definition only tells us the conditions an assignment of probabilities must satisfy; it doesn't tell us how to assign specific propabilities to events.

Probability measures satisfy some basic properties:

Proposition 2 (Basic properties of probability measure)**.** *If $\mathbb{P}$ is a probability measure, then the following properties hold:*

(i.) *(The null event has probability zero)* $\mathbb{P}[\emptyset] = 0$.

(ii.) *(Finite additivity) Let $\{E_1, \ldots, E_n\}$ be a finite sequence of events. If the sequence is pairwise disjoint, then*
$$\mathbb{P}[E_1 \cup E_2 \cup \ldots \cup E_n] = \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots + \mathbb{P}[E_n].$$

(iii.) *("With probability one, an event E either does occur or doesn't")* $\mathbb{P}[E^c] = 1 - \mathbb{P}[E]$.

(iv.) *(Excision Property) If A, B are events and $A \subseteq B$, then*
$$\mathbb{P}[B \backslash A] = \mathbb{P}[B] - \mathbb{P}[A].$$

(v.) *("The particular is less likely than the general") If A, B are events and $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.*

(vi.) *("Probabilities are between 0 and 1") For any event E, $\mathbb{P}[E] \in [0, 1]$.*

2 2026-01-14 | Week 01 | Lecture 02

The topic of this lecture: *independent events, conditional probabilities, random variables*

2.1 Independent events and conditional probabilities

This section is based on section 2.7 in the textbook.

Definition 3 (Independence)**.** Two events A and B are said to be ***independent*** if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Otherwise, the events are said to be dependent.

Definition 4 (Conditional probability)**.** Let A, B be events, and assume that $\mathbb{P}[B] > 0$. Then the ***conditional probability of A, given B***, denoted $\mathbb{P}[A \mid B]$, is given by the formula

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Interpretation: $\mathbb{P}[A \mid B]$ is the probability of A when we know that event B happened.

Definition 5. We say that there exists a ***positive relationship*** between events A and B if

$$\mathbb{P}[A \mid B] > \mathbb{P}[A],$$

and a ***negative relationship*** if

$$\mathbb{P}[A \mid B] < \mathbb{P}[A].$$

Remark 6. Note the the conditions of Definition 5 are symmetric in the sense that

$$\mathbb{P}[A \mid B] > \mathbb{P}[A] \iff \mathbb{P}[B \mid A] > \mathbb{P}[B],$$

provided that both A and B have positive probability.

Example 7. Roll a 6-sided dice. Let A be the event that a '2' was rolled, and B be the event that an even number was observed.

- The unconditional probability: $\mathbb{P}[A] = 1/6$.
- The conditional probability: $\mathbb{P}[A \mid B] = 1/3$.

Since $\frac{1}{3} > \frac{1}{6}$, we conclude there is a positive relationship between rolling a '2' and rolling an even number.

End of Example 7. □

The notion of independence formalizes the idea of "no relationship".

Proposition 8. *If A, B are events with positive probabilities then the following are equivalent:*

(i.) *A and B are independent.*

(ii.) $\mathbb{P}[A \mid B] = \mathbb{P}[A]$ *and* $\mathbb{P}[B \mid A] = \mathbb{P}[B]$.

In words, independence means that the probabilities of each event are unaffected by whether or not the other event occurs. Proposition 8 simply formalizes this idea using conditional probabilities.

2.2 Random variables

Based on Sections 2.11, 4.2 in the textbook

Definition 9 (Random variable)**.** A ***random variable*** (or ***rv***) is a real-valued function whose domain is a sample space.

The value of a random variable is thought of as varying depending on the outcome of the experiment (the sample point). Random variables are usually denoted with capital letters, like X, Y, Z.

Example 10 (Sum 2d4)**.** Roll a 4-sided dice twice (this is the **experiment**). There are 16 possible **outcomes**. The **sample space** is

$$S = \{(x, y) : x, y \in \{1, 2, 3, 4\}\}.$$

Let X be the sum of the two rolls. We can represent X by the following table:

		Dice 2			
		1	2	3	4
Dice 1	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8

Events are often defined using preimages of random variables. Most interesting take the form $[X \in B]$, where X is a random variable and $B \subseteq \mathbb{R}$. For example, the event that $X = 6$ is:

$$\begin{aligned}[X = 6] &= \{\omega \in S : X(\omega) = 6\} \\ &= \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.\end{aligned}$$

The textbook uses the notation $\{X = 6\}$ instead of $[X = 6]$.

Here's another example of an event. Let $E = \{2, 4, 6, 8\}$. Then

$$\begin{aligned}[X \text{ is even}] &= [X \in E] \\ &= \{\omega \in S : X(\omega) \in E\} \\ &= \{(1, 1), (1, 3), (2, 2), (2, 4), (3, 1), (3, 3), (4, 2), (4, 4)\}.\end{aligned}$$

When writing random variables, we usually suppress the arguments, e.g., writing X rather than $X(\omega)$.

End of Example 10. □

3 2026-01-16 | Week 01 | Lecture 03

3.1 Random variables

3.1.1 Discrete vs continuous

Definition 11 (Discrete random variable)**.** We say that a random variable X is a ***discrete random variable*** if it can assume only a finite or countably infinite number of distinct values.

Definition 12 (Probability mass function, pmf)**.** Let X be a discrete random variable. The ***probability mass function*** (or ***pmf***) of X is the function

$$p(x) = \mathbb{P}[X = x],$$

defined for every $x \in \mathbb{R}$.

Example 13. The pmf of X in Example 10 is

$$p(2) = 1/16 \quad p(3) = 2/16, \quad p(4) = 3/16, \quad p(5) = 4/16, \quad p(6) = 3/16, \quad p(7) = 2/16, \quad p(8) = 1/16$$

and $p(x) = 0$ for all other $x \in \mathbb{R}$.

End of Example 13. □

Definition 14 (Distribution function - section 4.2)**.** Let X be any random variable. The ***cumulative distribution function*** (or ***cdf***) of X is the function

$$F(x) = \mathbb{P}[X \leq x],$$

defined for all $x \in \mathbb{R}$.

Remark 15. The domain of a cdf is always $\mathbb{R}$, and it is always a nondecreasing function with $F(-\infty) = 0$ and $F(+\infty) = 1$. The cdf of a discrete random variable is always a step function.

Definition 16 (Continuous rv)**.** Let Y be a random variable with distribution function F. We say that Y is a ***continuous random variable*** if there exists a nonnegative function f such that

$$F(y) = \int_{-\infty}^{y} f(t)dt \tag{1}$$

for all $y \in \mathbb{R}$. The function f is called the ***probability density function*** (or ***pdf***) of Y.

Remark 17. For continuous random variables, the distribution function F is always continuous. Moreover, for a continuous random variable Y, $\mathbb{P}[Y = b] = 0$ for all $b \in \mathbb{R}$.

Theorem 18 (Theorem 4.3 in textbook)**.** *If Y is a continuous random variable with pdf f, then*

$$\mathbb{P}[a \leq Y \leq b] = \int_{a}^{b} f(t)dt$$

for all $-\infty \leq a \leq b \leq +\infty$.

3.1.2 Expected value

Definition 19 (Expectation of a continuous random variable)**.** If Y is a random variable with pdf f, then the ***expected value*** of Y, denoted $\mathbb{E}[Y]$, is the quantity

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)dy,$$

provided that $\int_{-\infty}^{\infty} |y| f(y)dy < \infty$.

Remark 20. $\mathbb{E}[Y]$ is the long-run average of Y, if we were to repeat the experiment many times.

The next theorem is called the *Law of the unconscious statistician (LOTUS).*

Theorem 21 (LOTUS - single variable case)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a function.*

(i.) *If X has pmf p, then*

$$\mathbb{E}[g(X)] = \sum_{\substack{x \in \mathbb{R} \\ p(x) > 0}} g(x)p(x).$$

(ii.) *If Y has pdf f, then*

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy.$$

Remark 22. Often we wish to compute probabilities of functions of multiple random variables, for example:

- What is the probability that $\frac{X_1+\ldots+X_n}{n} \in (0,1)$? Here, the function is $g(x_1, \ldots, x_n) = \frac{x_1+\ldots+x_n}{n}$.
- What is the probability that $\max(X, Y) \leq 10$? Here the function is $g(x, y) = \max(x, y)$.
- Suppose we roll two dice and take the maximum. What is the expected value? In this case, our dice rolls are X, Y and we want to compute $\mathbb{E}[g(X, Y)]$, where $g(x, y) = \max(x, y)$.

To answer these sorts of questions, we need the notion of a "joint distribution".

3.1.3 Joint distributions

This subsection is based on section 5.4 in the textbook. Everything in this section generalizes naturally to n variables, but the results are simpler to state for just 2 random variables.

Definition 23 (Joint pmf)**.** Let X_1 and X_2 be discrete random variables. The ***joint probability mass function*** for X_1 and X_2 is the function

$$p(x_1, x_2) = \mathbb{P}[X_1 = x_1, X_2 = x_2],$$

defined for all $x_1, x_2 \in \mathbb{R}$.

Definition 24 (Joint pdf)**.** Let Y_1 and Y_2 be continuous random variables. We say that Y_1 and Y_2 are ***jointly continuous*** if there exists a function $f : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}[Y_1 \leq y_1, Y_2 \leq y_2] = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2)dt_2dt_1.$$

for all $y_1, y_2 \in \mathbb{R}$. The function f is called the ***joint probability density function*** for Y_1 and Y_2.

Theorem 25 (LOTUS - multivariable case)**.** *Let $g : \mathbb{R}^2 \to \mathbb{R}$.*

- *If X_1, X_2 have joint pmf $p(x_1, x_2)$, then*

$$\mathbb{E}[g(X_1, X_2)] = \sum_{\substack{(x_1,x_2) \in \mathbb{R}^2: \\ p(x_1,x_2) > 0}} g(x_1, x_2)p(x_1, x_2).$$

- *If Y_1, Y_2 are jointly continuous random variables with joint pdf $f(y_1, y_2)$, then*

$$\mathbb{E}[g(Y_1, Y_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2)f(y_1, y_2)dy_1dy_2.$$

Remark 26. Theorem 25 generalizes to n variables. It gives us a way to answer questions like the third question posed in Remark 22.

4 2026-01-21 | Week 02 | Lecture 04

> ***Question for this lecture:*** *what does it mean for random variables to be independent, and what does it buy us?*

4.1 Independence of random variables

4.1.1 Definition and characterization

The aim of this section is to define what it means for random variables to be independent.

Definition 27. We say that the random variables $X_1, X_2, \ldots, X_n$ are ***independent*** if the following holds for all possible values $x_1, \ldots, x_n$ in their range:

$$\mathbb{P}[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1]\,\mathbb{P}[X_2 \leq x_2] \cdots \mathbb{P}[X_n \leq x_n].$$

Theorem 28 (Factorization theorem – Theorem 5.4 in textbook)**.** *For discrete/continuous random variables, independence is equivalent to factorizability of the joint pmf/pdf. More formally, we have:*

- ***Discrete case:*** *Let $X_1, \ldots, X_n$ be discrete random variables. Then $X_1, \ldots, X_n$ are independent if and only if*
$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \mathbb{P}[X_1 = x_1]\,\mathbb{P}[X_2 = x_2] \cdots \mathbb{P}[X_n = x_n]$$
for all $x_1, \ldots, x_n \in \mathbb{R}$.
- ***Continuous case:*** *Let $Y_1, \ldots, Y_n$ be continuous random variables with pdfs $f_1, \ldots, f_n$, and joint pdf $f : \mathbb{R}^n \to \mathbb{R}$. Then $Y_1, \ldots, Y_n$ are independent if and only if*
$$f(y_1, \ldots, y_n) = f_1(y_1) f_2(y_2) \cdots f_n(y_n)$$
for all $y_1, \ldots, y_n \in \mathbb{R}$.

Remark 29. In this course, we will frequently work with "samples" of n independent random variables $X_1, \ldots, X_n$. The intuition about independence for n variables is that observing any number of them doesn't give you any information about the others.

4.1.2 Independence and expectations/variances

Independence splits the expectation of a product into a product of expectations.

Theorem 30. *If $X_1, X_2, \ldots X_n$, are independent, then*

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mathbb{E}[X_1]\,\mathbb{E}[X_2] \cdots \mathbb{E}[X_n],$$

provided that the expectations exist.

Proof. This can be proven for continuous/discrete random variables by an application of Theorem 25. □

Definition 31 (Variance of a random variable)**.** Let X be a random variable. The ***variance*** of X, denoted $\mathrm{Var}(X)$, is the quantity

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mu)^2\right].$$

where $\mu = \mathbb{E}[X]$. The positive square root of the variance is the ***standard deviation*** of X.

Remark 32. Variance often denoted by σ^2. The textbook uses the notation $V(X)$ instead of $\mathrm{Var}(X)$.

Theorem 33. *Let X be a random variable and a, b be scalars. Then*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

If X and Y are independent random variables then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof. Follows by direct computation using the definition of variance. □

The second part of Theorem 33 says that for independent random variables, variance is additive.

4.1.3 Independence and long-term behavior: the weak law of large numbers

Combining many independent sources of randomness tends to produce predictible phenmonena. Here we introduce one example of that.

We will need the following theorem, which says that if the average male height is 5 feet tall, then no more than 10% of men can be more than 50 feet tall.

Theorem 34 (Markov's inequality)**.** *Let X be a nonnegative random variable and let $a > 0$. Then*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a} \tag{2}$$

Proof. Let $E = [X \geq a]$. Then $X \cdot \mathbf{1}_E \leq X$, so

$$\begin{aligned}\mathbb{E}[X] &\geq \mathbb{E}[X \cdot \mathbf{1}_E] \\ &\geq \mathbb{E}[a\mathbf{1}_E] \\ &\geq a\mathbb{E}[\mathbf{1}_E] \\ &= a\mathbb{P}[E] \\ &= a\mathbb{P}[X \geq a].\end{aligned}$$

Dividing by a implies Eq. (2). □

Theorem 35 (Weak Law of Large Numbers)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables. Assume that for all n, $\mu = \mathbb{E}[X_n]$ and $\mathrm{Var}(X_n) \leq B$ for some fixed bound $B < \infty$. Let $S_n = X_1 + \ldots + X_n$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] \to 0 \text{ as } n \to \infty.$$

Proof. First observe that by independence,

$$\mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) \leq nB. \tag{3}$$

Next, let $\epsilon > 0$ be arbitrary.

$$\begin{aligned}\mathbb{P}\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= \mathbb{P}[|S_n - n\mu| > n\epsilon] \\ &= \mathbb{P}\left[(S_n - n\mu)^2 > (n\epsilon)^2\right] \\ &\leq \frac{\mathbb{E}\left[(S_n - n\mu)^2\right]}{n^2\epsilon^2} && \text{by Theorem 34} \\ &= \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} && \text{by definition of variance} \\ &\leq \frac{B}{\epsilon^2 n} && \text{by Eq. (3).}\end{aligned}$$

The right hand side tends to zero as $n \to \infty$, proving the theorem. □

4.2 Equality in distribution

Definition 36 (Identically distributed)**.** Two random variables X and Y are ***identically distributed*** if they have the same cdf.

Remark 37. Identically distributed random variables are said to "have the same distribution". For discrete/continuous random variables, this is equivalent to them having the same pmf/pdf.

The next example shows that identically distributed random variables need not be equal as functions.

Example 38. Flip a coin. Define two random variables

$$X = \begin{cases} 1 & : \text{ coin is heads} \\ 0 & : \text{ coin is tails} \end{cases}$$

$$Y = \begin{cases} 0 & : \text{ coin is heads} \\ 1 & : \text{ coin is tails} \end{cases}$$

Then X and Y are identically distributed, since they have the same pmf:

$$\mathbb{P}[X = 1] = \mathbb{P}[Y = 1] = \frac{1}{2} \quad \text{and} \quad \mathbb{P}[X = 0] = \mathbb{P}[Y = 0] = \frac{1}{2}.$$

But of course X and Y are never equal, i.e., $\mathbb{P}[X \neq Y] = 1$.

End of Example 38. □

4.3 Populations and samples

For discussions of "random sample", see sections 2.12 and 6.1 in the textbook.

Here we introduce a conceptual framework for mathematical statistics.

A ***population*** is a large body of data that is the target of our interest. The subset collected from it is our ***sample***.

Example 39 (Populations)**.** A population can be real or theoretical. Here are some examples

- The set of people in Hawaii (real, finite)
- The set of voters in the 2026 Midterm elections (hypothetical, finite)
- The decimal expansion of π (countably infinite)
- The infinitely many observations that could be made during a laboratory experiment if the experiment were repeated over and over again (hypothetical)
- The lifetimes of light bulbs produced by a factory

Importantly, a population can also be a *probability distribution*, specified by a pdf, pmf, or cdf

- Observations made from an exponential distribution with mean $\lambda > 0$ (i.e., the distribution with pdf $f(x) = \lambda e^{-\lambda x} \mathbf{1}_{[x>0]}$.)

End of Example 39. □

The simpleset sampling procedure is called simple random sampling.

Definition 40 (Simple random sampling)**.** Let N and n denote the numbers of elements in the population and sample, respectively. If the sampling is conducted in such a way that each of the $\binom{N}{n}$ samples has an equal probability of being selected, the sampling is called ***simple random sampling***, and the result is a ***simple random sample***.

More commonly in mathematical statistics, we think of the population as a distribution.

Definition 41 (Random sample from a distribution)**.** Consider a given probability distribution on $\mathbb{R}$ that can be represented by a pdf or pmf f. We say that the random variables $X_1, \ldots, X_n$ form a ***random sample*** from this distribution if these random variables are independent and the distribution of each is given by f. Such random variables are also said to be ***independent and identically distributed*** (iid). The number of random variables n is the ***sample size***.

The objective of statistics to is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference.

4.4 What is a statistic?

Section 7.1 in the textbook.

Definition 42 (Statistic)**.** A ***statistic*** is a function of the observable random variables in a sample and known constants.

In other words, if $X_1, \ldots, X_n$ is a random sample and $T : \mathbb{R}^n \to \mathbb{R}$ is a function, then the random variable

$$Y = T(X_1, \ldots, X_n)$$

is a statistic.[2] The probability distribution of such a statistic Y is called its ***sampling distribution***.

Often, we have some quantity of interest, called a ***target parameter***, and we want a single "best guess" of some quantity of interest. When this is the case, the we call a statistic an estimator:

Definition 43 (Estimator)**.** An ***estimator*** is a statistic, that is a function $T(X_1, \ldots, X_n)$ of a sample, that is used to approximate a target parameter. An ***estimate*** is the realized value of an estimator (e.g., a number) that is obtained when the sample is actually taken.

In words, an estimator is a rule, often expressed as a formula, that tells us how to calculate the value of an estimate based on the measurements contained in a sample. Here are two important examples of estimators.

Definition 44 (Sample Mean, Sample Variance)**.** Let $X_1, \ldots, X_n$ be a random sample. The ***sample mean***, denoted $\overline{X}$, is the statistic

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The ***sample variance***, denoted S^2, is the statistic

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The target parameters these are used to estimate are the mean and variance of the population.

The following examples illustrates some of the terminology we've introduced.

Example 45. Suppose the lifetime of each light bulb produced in a certain factory is distributed according to the following pdf

$$f(y) = \lambda e^{-\lambda y}\mathbf{1}_{[y>0]}.$$

for some parameter $\lambda > 0$, where y is measured in minutes. Here, the **population** is the distribution of lifetimes. Suppose we do not know λ, and we wish to estimate it. Hence λ is the **target parameter**.

A random sample $Y_1, \ldots, Y_n$ is taken. (So the Y_i's are independent and each as pdf f.)

Claim 1: The **joint pdf** of $Y_1, \ldots, Y_n$ is the function

$$g(y_1, \ldots, y_n) = \lambda^n e^{-\lambda(y_1+\ldots+y_n)}\mathbf{1}_{[\min(y_1,\ldots,y_n)>0]},$$

defined for all $y_1, \ldots, y_n \in \mathbb{R}$

Proof of Claim 1. Using the fact that $Y_1, \ldots, Y_n$ in independent continuous random variables,

$$\begin{aligned} g(y_1, \ldots, y_n) &= f(y_1)\cdots f(y_n) && \text{by Theorem 28} \\ &= \prod_{i=1}^{n}\left(\lambda e^{-\lambda y_i}\mathbf{1}_{[y_i>0]}\right) \\ &= \lambda^n e^{-\lambda(y_1+\ldots+y_n)}\left(\prod_{i=1}^{n}\mathbf{1}_{[y_i>0]}\right) \\ &= \lambda^n e^{-\lambda(y_1+\ldots+y_n)}\mathbf{1}_{[\min(y_1,\ldots,y_n)>0]}. \end{aligned}$$

[2] A statistic does not need to be real-valued, but we will focus on the case when it is.

□ Claim

Claim 2: If Y has pdf f, then $\mathbb{E}[Y] = \frac{1}{\lambda}$.

Proof of Claim 2.

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty} y\lambda e^{-\lambda y}\mathbf{1}_{[y>0]}dy \\
&= \int_{0}^{\infty} y\lambda e^{-\lambda y}dy \\
&= \int_{0}^{\infty} e^{-\lambda y}dy && \text{by integration by parts} \\
&= \frac{1}{\lambda}.
\end{aligned}$$

□ Claim

By Theorem 35 (The Weak Law of Large Numbers), the **statistic**

$$\overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}$$

is likely to be close to $\mathbb{E}[Y] = 1/\lambda$ when n is large. Therefore a reasonable **estimator** for λ is

$$\frac{1}{\overline{Y}} = \frac{n}{Y_1 + \ldots + Y_n}.$$

If $n = 10$ and the observed values of $Y_1, \ldots, Y_{10}$ are

$$2,\ 0.8,\ 0.1,\ 2.3,\ 3.2,\ 5.4,\ 2,\ 0.4,\ 2.8,\ \text{ and } 3.4$$

then $\overline{Y} = 2.24$, and hence our **estimate** for λ would be $\frac{1}{2.24} \approx 0.45$.

End of Example 45. □

Example 46 (Example 7.1 in textbook)**.** Roll a dice three times. Let Y_1, Y_2, Y_3 be the values of the rolls. The average number observed in this sample of size 3 is

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3}{3}.$$

This is the **sample mean**, and is the first example of a **statistic** that we'll see. What the mean $\mu_{\overline{Y}}$ and standard deviation $\sigma_{\overline{Y}}$ of the random variable $\overline{Y}$?

First we compute the mean and variance of a single Y_i.

Claim 1: $\mathbb{E}[Y_i] = 3.5$ and $\mathrm{Var}(Y_i) = 35/12 \approx 2.9167$.

Proof of Claim 1. Using the definition of expected value,

$$\mathbb{E}[Y_i] = \sum_{k=1}^{6} k\mathbb{P}[Y_i = k] = \sum_{k=1}^{6} k\frac{1}{6} = 3.5.$$

Next we compute the variance of Y_i:

$$\begin{aligned}
\mathrm{Var}(Y_i) &= \mathbb{E}\left[(Y_i - 3.5)^2\right] && \text{Definition 31} \\
&= \sum_{k=1}^{6}(k-3.5)^2\mathbb{P}[Y_i = k] && \text{By Theorem 21 (LOTUS) using } g(x) = (x-3.5)^2 \\
&= \sum_{k=1}^{6}(k-3.5)^2\frac{1}{6} && \text{since } \mathbb{P}[Y_i = k] = 1/6 \text{ for all } k \in [6] \\
&= 35/12.
\end{aligned}$$

□ Claim

Claim 2: $\mathbb{E}\left[\overline{Y}\right] = 3.5$

Proof of Claim 2. We have:

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}\right] &= \mathbb{E}\left[\frac{Y_1 + Y_2 + Y_3}{3}\right] && \text{by definition of } \overline{Y} \\
&= \frac{1}{3}\left(\mathbb{E}\left[Y_1\right] + \mathbb{E}\left[Y_2\right] + \mathbb{E}\left[Y_3\right]\right) && \text{by linearity of expectation} \\
&= \frac{1}{3}\left(3.5 + 3.5 + 3.5\right) && \text{by Claim 1} \\
&= 3.5.
\end{aligned}$$

□ Claim

Claim 3: $\text{Var}(\overline{Y}) = .9722$

Proof of Claim 3.

$$\begin{aligned}
\text{Var}(\overline{Y}) &= \text{Var}\left(\frac{Y_1 + Y_2 + Y_3}{3}\right) \\
&= \frac{1}{9}\text{Var}\left(Y_1 + Y_2 + Y_3\right) && \text{by Theorem 33} \\
&= \frac{1}{9}\left(\text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3)\right) && \text{by Theorem 33} \\
&= \frac{1}{9}\left(2.9167 + 2.9167 + 2.9167\right) && \text{by Claim 1} \\
&= \frac{2.9167}{3} && (*) \\
&= .9722
\end{aligned}$$

□ Claim

Since the standard deviation is the square root of the variance, the standard deviation is $\sqrt{.9722} = .9860$ by Claim 3.

Remark 47. In equation $(*)$, we had $\text{Var}(\overline{Y}) = \frac{2.9167}{3}$. Notice how, if we computed the variance of four or five dice rolls, we'd get

$$\text{Var}(\overline{Y}) = \frac{2.9167}{4} \quad \text{or} \quad \text{Var}(\overline{Y}) = \frac{2.9167}{5}.$$

In general, for n dice rolls, we'd get

$$\text{Var}(\overline{Y}) = \frac{2.9167}{n}.$$

This quantity tends to zero as $n \to \infty$. The variance of a sample mean decreases as n grows.

End of Example 47. □

Begin section 7.2

Definition 48. A random variable Y is said to be ***normally distributed*** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ if it has pdf

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},$$

defined for all $y \in \mathbb{R}$. We abbreviate this by writing $Y \sim \mathcal{N}(\mu, \sigma^2)$. If $Y \sim \mathcal{N}(0, 1)$, we say Y has ***standard normal distribution***.

Theorem 49. *Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$.

Proof sketch. The proof follows from three facts:

- Any linear combination of normal random variables is a normal random variable (this can be proved using moment generating functions).
- That $\mathbb{E}\left[\overline{Y}\right] = \mu$ follows by linearity of expectation.
- The variance calculation follows by Theorem 33.

□

Remark 50. The random variable $Z = \frac{\overline{Y}-\mu_{\overline{Y}}}{\sigma_{\overline{Y}}} = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

Example 51 (Example 7.2 in textbook)**.** A bottling machine discharages an average of μ ounces per bottle. It has been observed that the amount of liquid discharged is normally distributed with standard deviation $\sigma = 1$ per ounce.

A sample of $n = 9$ filled bottles is randomly selected, and is measured.

Find the probability that the sample mean will be within .3 ounces of the true mean μ.

Solution: Let $Y_1, \ldots, Y_9$ be the ounces in each of the bottles. Then $Y_i \sim \mathcal{N}(\mu, 1)$ for $i \in [9]$. By Theorem 49, $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$, and $\sigma^2/n = 1/9$. We want to find

$$\begin{aligned}\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le .3\right] &= \mathbb{P}\left[-.3 \le \overline{Y}-\mu \le .3\right] \\ &= \mathbb{P}\left[-\frac{.3}{\sigma/\sqrt{n}} \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le \frac{.3}{\sigma/\sqrt{n}}\right]. \end{aligned} \tag{4}$$

By Remark 50, $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, and plugging in $\sigma = 1$ and $n = 9$ and simplifying gives

$$\begin{aligned}\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le .3\right] &= \mathbb{P}\left[-.9 \le Z \le .9\right] \\ &= 1 - 2\mathbb{P}\left[Z > .9\right] \\ &= .6318.\end{aligned}$$

End of Example 51. □

Example 52 (Example 7.3 in textbook)**.** How many observations need to be included in the sample to ensure that $\overline{Y}$ is within .3 ounces of μ with probability .95?

Now we want

$$\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le .3\right] = .95.$$

By Eq. (4),

$$\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le .3\right] = \mathbb{P}\left[-\frac{.3}{\sigma/\sqrt{n}} \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le \frac{.3}{\sigma/\sqrt{n}}\right].$$

Recalling that $\sigma = 1$ and $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, we have

$$\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le .3\right] = \mathbb{P}\left[.3\sqrt{n} \le Z \le .3\sqrt{n}\right].$$

So we need to find n such that $\mathbb{P}\left[.3\sqrt{n} \le Z \le .3\sqrt{n}\right] = .95$.

It is a well-known fact that

$$\mathbb{P}\left[-1.96 \le Z \le 1.96\right] = .95.$$

So we need $.3\sqrt{n} = 1.96$, or equivalently, $n = \left(\frac{1.96}{.3}\right)^2 = 42.68$.

So we need at least 43 samples, since 42 is not quite enough.

End of Example 52. □

4.5 Chi-squired distributions

Define chi-squared distribution, state theorem 7.2.