Lecture Notes for Math 472: Statistical Inference

Last updated: February 27, 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 More realistic course outline

1. **Weeks 1-4.** General framework of probability and statistics

   sample space, probability measure, independent events, random variable, expectation, variance, joint distribution, independent random variables, law of large numbers

   sample-population framework, statistics, sample mean, sample variance, Gaussian distribution, standardization, t-distribution, chi-squared distribution, studentization, central limit theorem, point estimation, mean square error

2. **Weeks 5-7.** Interval estimation

   k-standard error bounds (8.4), confidence intervals (8.5), large-sample confidence intervals (8.6), comparing two populations (8.6), sample complexity (8.7), small-sample confidence intervals (8.8), confidence intervals for variance (8.9)

3. **Weeks 8-9.** Properties and methods of point estimation

   relative efficiency (9.2), consistency (9.3), sufficiency (9.4), likelihood (9.4) factorization theorem (9.4), Rao-Blackwell Theorem (9.5), MVUE (9.5), method of moments (9.6), method of maximum likelihood (9.7)

4. **Weeks 10-12.** Hypothesis Testing

   hypothesis testing (10.2), common large-sample tests (10.3), type-II error probabilties (10.4), connection to confidence intervals (10.5), power (10.10), Neyman-Pearson lemma (10.10), likelihood ratio test (10.11), chi-squared test (14.2-14.5)

   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

   (11.2-11.5, 11.8): regression, bivariate normal distributions, method of least squares, ANOVA? (chapter 13)

2 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.

2.1 What is probability?

2.1.1 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.

2.1.2 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, $\mathbb{P}[E] \in [0, 1]$.*

3 2026-01-14 | Week 01 | Lecture 02

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

3.1 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.

3.2 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. □

4 2026-01-16 | Week 01 | Lecture 03

4.1 Random variables

4.1.1 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 $-\infty \leq a \leq b \leq +\infty$.

4.1.2 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".

4.1.3 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.

5 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?*

5.1 Independence of random variables

5.1.1 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.

5.1.2 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 and Theorem 28. □

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

Theorem 33 (Properties of Variance)**.** *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.

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

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

Theorem 34 (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 $\epsilon > 0$,*

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

6 2026-01-23 | Week 02 | Lecture 05

6.1 Proof of the weak law of large numbers

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 35 (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 $A = [X \geq a]$. Then $X \cdot \mathbf{1}_A \leq X$, so

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

Dividing by a implies Eq. (2). □

Theorem 36 (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.

$$\begin{aligned}\mathbb{P}\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= \mathbb{P}[|S_n - n\mu| > n\epsilon] \\ &= \mathbb{P}\left[(S_n - n\mu)^2 > (n\epsilon)^2\right] \\ &\leq \frac{\mathbb{E}\left[(S_n - n\mu)^2\right]}{n^2\epsilon^2} && \text{by Theorem 35} \\ &= \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} && \text{by definition of variance} \\ &\leq \frac{B}{\epsilon^2 n} && \text{by Eq. (3).}\end{aligned}$$

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

6.2 Equality in distribution

Definition 37 (Identically distributed)**.** Two random variables X and Y are ***identically distributed*** if they have the same cdf.

Remark 38. 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 39. 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 39. □

6.3 Populations and samples

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

- The set of people in Hawai'i (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 40. □

7 2026-01-26 | Week 03 | Lecture 06

7.1 Sampling

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

The simpleset sampling procedure is called simple random sampling.

Definition 41 (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 42 (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.

7.2 What is a statistic?

Section 7.1 in the textbook.

Definition 43 (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 44 (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 45 (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.

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

Example 46 (Estimating average lightbulb lifetime)**.** 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 either the set of lightbulbs produced by the factory, which we idealize as the probability distribution given by f. 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.)

In this case f is an exponential distribution with rate $\lambda > 0$. The exponential distribution is used to modeling lifetimes of systems that are not subject to degredation (e.g., lifetimes of light bulbs, not diesel engines) as well as waiting times (e.g., the length of time intervals between cars passing by a tree on an idyllic country road). In the road example, if the rate of cars passing is 2 cars per minute, then on average you have to wait only 1/2 of a minute for a car to pass. That is, the *rate* and *expected values* are reciprocals of each other. Let's prove this fact in the following claim:

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

Proof of Claim 1. By definition of expected value for a continuous random variable,

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

□ Claim

End of Example 46. □

8 2026-01-28 | Week 03 | Lecture 07

8.1 Basic examples of statistics and estimators

This lecture is based on section 7.1 in the textbook.

Example 47 (Example 46 continued)**.** We have a random sample of $Y_1, \ldots, Y_n$ from the distribution with pdf

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

where λ is a target parameter.

In the setting of this example, we interpret the random variables $Y_1, \ldots, Y_n$ as lifetimes (in, say, years) of lightbulbs produced by a certain factory.

In the last lecture, we used integration by parts to show that if Y has pdf f, then

$$\mathbb{E}[Y] = \frac{1}{\lambda}.$$

Consider the **statistic**

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

By Theorem 36 (The Weak Law of Large Numbers), $\overline{Y}$ is likely to be close to $\mathbb{E}[Y] = 1/\lambda$ when n is large. Therefore $\frac{1}{\overline{Y}}$ is likely to be close to λ. 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 47. □

Example 48 (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**, it is function of the sample (and the known constant n) and is therefore a **statistic**.

Question: 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. Fix $i \in [3]$. For simplicity, let $Y = Y_i$. Using the definition of expected value,

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

Next we compute the variance of Y:

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

□ Claim

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

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: $\sigma^2_{\overline{Y}} = \mathrm{Var}(\overline{Y}) = .9722$

Proof of Claim 3.

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

and this is approximately .97224. □ Claim

Since the standard deviation is the square root of the variance, the standard deviation is

$$\sigma_{\overline{Y}} = \sqrt{.9722} = .9860$$

Claim 3.

Remark 49. In equation $(*)$, we had $\mathrm{Var}(\overline{Y}) = \frac{2.9167}{3}$. Notice how, if we computed the variance of four or five dice rolls, we'd get

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

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

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

This quantity tends to zero as $n \to \infty$. In words, as the number of samples n grows, the variance of the sample mean decreases.

End of Example 49. □

9 2026-01-30 | Week 03 | Lecture 08

9.1 The Gaussian distribution

Begin section 7.2

Definition 50 (Normal / Gaussian Distribution)**.** 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***.

Remark 51. Normal random variables/distributions are also called ***Gaussian*** random variables/distributions.

Example 52. Normal distributions are the classic "bell curve" distributions. Whenever a random quantity is the result of adding up many small, independent random effects, we expect it to be approximately normally distributed.

- A classic example is human height, which is influenced by the combined effects of thousands of genes together with environmental factors.
- Another example is repeated measurements of the same physical quantity (e.g., if we all went out and measured the temperature outside tomorrow at midday), this is because the total error is the sum of many small independent sources of noise (e.g., instrument precision, time, location, shade, wind, etc).

End of Example 52. □

Remark 53 (Standardization)**.** If $Y \sim \mathcal{N}(\mu, \sigma^2)$ then the random variable

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

is a standard normal random variable. The transformation

$$X \mapsto \frac{X - \mu}{\sigma}$$

in which we subtract off the mean of X and then divide by its standard deviation is called **standardization**.

Remark 54 (The Empirical Rule)**.** The **Empirical Rule** (also called the **68-95-99.7 rule**; see textbook section 1.3) is summarized by the following figure, for normal distributions:

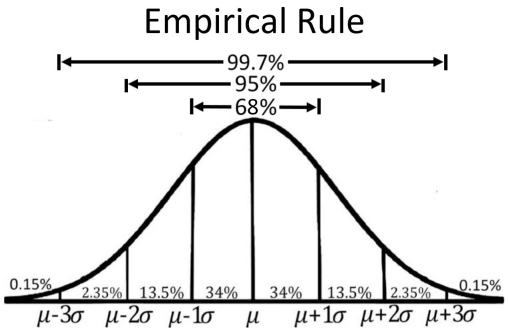


Image source: https://andymath.com/wp-content/uploads/2019/12/empirical-rule-normdist2.jpg

The next theorem says that if we have a sample of $Y_1, \ldots, Y_n$ normal random variables, then their sample mean is normally distributed as well.

Theorem 55 (Mean and variance of a normal sample mean)**.** *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 independent 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 56. Applying the idea of standardization, observe that 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 57 (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.

Question: 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 55, $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$, 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 56, $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, 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 57. □

Example 58 (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

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

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 58. □

10 2026-02-02 | Week 04 | Lecture 09

This lecture is based on Section 7.2 in the textbook. I will be skipping the part about F statistics for now.

10.1 Motivation: "Approximate" standardization

In Example 57, we had a random sample $Y_1, \ldots, Y_n$ of normally distributed random variables, and we computed $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ by rewriting the even $\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ using the standardization trick:

$$\begin{aligned}\left[\left|\overline{Y} - \mu\right| \leq .3\right] &= \left[-.3 \leq \overline{Y} - \mu \leq .3\right] \\ &= \left[\frac{-.3}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{.3}{\sigma/\sqrt{n}}\right] \\ &= \left[\frac{-.3}{\sigma/\sqrt{n}} \leq Z \leq \frac{.3}{\sigma/\sqrt{n}}\right]\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. Hence,

$$\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right] = \left[\frac{-.3}{\sigma/\sqrt{n}} \leq \underbrace{\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}}_{Z} \leq \frac{.3}{\sigma/\sqrt{n}}\right]. \tag{5}$$

In other words, we used standardization (see Remark 53) to rewrite the probability into something we could compute (or look up).

The problem with this approach is that the example is contrived and unrealistic because, in the problem statement, we assumed that we knew the true standard deviation to be $\sigma = 1$. In practice, this sort of thing is almost never known ahead of time, but instead must be estimated from the data.

A natural estimator of σ is S, the square root of the sample variance. This is because $S^2 \to \sigma^2$ as $n \to \infty$ by the law of large numbers, so S should be close to σ when n is large. In that case, we can **"approximately" standardize.** Replacing σ by S in the calculation above, we obtain

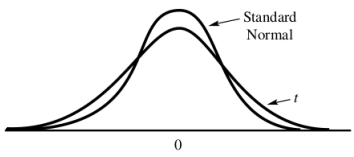
$$\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right] = \left[\frac{-.3}{S/\sqrt{n}} \leq \underbrace{\frac{\overline{Y} - \mu}{S/\sqrt{n}}}_{\text{call it } T} \leq \frac{.3}{S/\sqrt{n}}\right].$$

Here, the random variable T is not normally distributed, but it should be pretty close to $\mathcal{N}(0,1)$ when n is large. This is because $S \approx \sigma$ when n is large, so

$$T = \left(\frac{\overline{Y} - \mu}{S/\sqrt{n}}\right) \approx \left(\frac{\overline{Y} - \mu}{S/\sqrt{n}}\right) = Z.$$

Question: What is the distribution of T? Can we quantify "how close" it is to Z?
Short answer: Assuming a sample size of $n > 2$, T has a distribution called a t-distribution, which is well-known and widely used. It has expectation $\mathbb{E}[T] = 0$ and variance $\mathrm{Var}(T) = \frac{\nu}{\nu-2}$, where $\nu = n - 1$ is called the "degrees of freeom". It has a probability density function which is symmetric and bell-shaped, like the normal distribution, but the shape is a little more spread out than the standard normal:

FIGURE 7.3
A comparison of the standard normal and t density functions.



The reason for this greater spread is that $\mathrm{Var}(T) > 1$ for all $n > 2$, whereas $\mathrm{Var}(Z) = 1$.

Long answer: see the next two subsections.

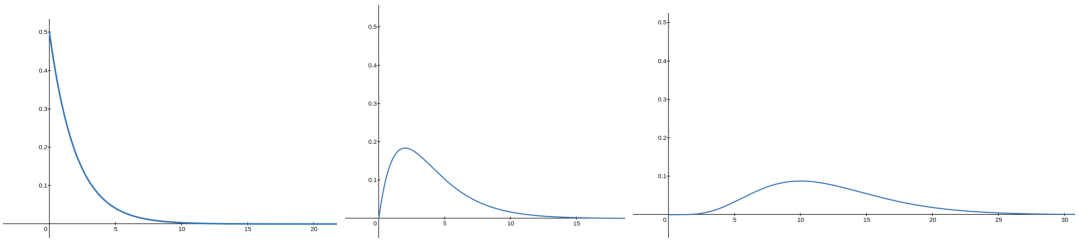
10.2 The chi-squared distribution

Definition 59 (Chi-Squared Distribution)**.** Let d be a positive integer. A nonnegative random variable X is said to have ***chi-squared distribution with parameter*** d if it has pdf

$$f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x/2} & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

The parameter d is usually called the ***degrees of freedom*** of X.

From left-to-right, here's what this looks like for $d = 2$, $d = 4$, and $d = 12$:



Proposition 60 (Properties of chi-squared distribution)**.** *A chi-squared random variable X with d degrees of freedom has*

$$\mathbb{E}[X] = d \quad \textit{and} \quad \mathrm{Var}(X) = 2d$$

Chi-squared random variables are important in statistics because they arise as squared normal random variables:

Theorem 61 (Chi-squared is sum of squared normals)**.** *Let $n \geq 1$. The following are equivalent*

(i.) *X has chi-squared distribution with n degrees of freedom*

(ii.) *$X \stackrel{d}{=} Z_1^2 + \ldots + Z_n^2$, where $Z_1, \ldots, Z_n$ are independent standard normal random variables*

A standard proof of Theorem 61 involves the use of moment generating functions. (See Example 6.11 in the textbook).

Remark 62. Taking $n = 1$, Theorem 61 implies that $Z \sim \mathcal{N}(0, 1)$ implies Z^2 is chi-squared with 1 degree of freedom.

Corollary 63 (Theorem 7.2 in textbook)**.** *If $Y_1, \ldots, Y_n$ is a sample of $\mathcal{N}(\mu, \sigma)$ random variables, then*

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

is a chi-squared random variable with $\mathrm{df} = n$.

Remark 64. Usually, we use a table or a computer to compute probabilites with chi-squared random variables. But the fact that we usually can't do exact computations by hand doesn't make chi-squared random variables any less important to statistics, as they show up in samples variances, linear regression, distance estimates (pythagorean theorem has squares), etc.

The main example we will need is the following

Theorem 65 (Theorem 7.3 in textbook)**.** *Let $Y_1, \ldots, Y_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then the random variable*

$$\frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $n - 1$ degrees of freedom. Also, the random variables $\overline{Y}$ and S^2 are independent.

See textbook for outline of the proof (for the case when $n = 2$). The independence of $\overline{Y}$ and S^2 is rather technical to prove, but simply says that the knowledge of the center of location of a normal random variable does not contribute to knowledge of its variability.

10.3 Definition of the Student t-distribution

Definition 66 (Student's t-distribution)**.** Let $Z \sim \mathcal{N}(0,1)$ and let χ^2_ν be an independent chi-squared distributed random variable with $\text{df} = \nu$. Then

$$T = \frac{Z}{\sqrt{\chi^2_\nu/\nu}}$$

is said to have a ***t-distribution***[3] with ν degrees of freedom.

Remark 67.

- To show that a random variable follows a T distribution, we must show that it can be written as a ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom.
- There are infinitely many t-distributions, one for every positive integral ν.
- As ν increases, the bell-curve distribution of T becomse less spread out, and approaches the standard normal curve as $\nu \to \infty$.

[3]This is also called "Student's t distribution". It was discovered by W. S. Gosset, an employee of the Guinness Brewery in Dublin, who wrote under the pseudonym "Student."

11 2026-02-04 | Week 04 | Lecture 10

11.1 Using the student t-distribution

This is based on section 7.2 of the textbook.

The following theorem illustrates an important example of the t-distribution, which arises when we "approximately standardize" a sample mean with the transformation

$$\overline{Y} \mapsto \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

(where we've used the sample standard deviation S in place of the unknown population paramter σ). This transformation, which involves dividing by S rather than σ is called ***studentization*** (as opposed to ***standardization*** when σ is used.)

Theorem 68 (Distribution of a studentized sample mean)**.** *Let $Y_1, \ldots, Y_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. The random variable*

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

follows a Student t-distribution with $n-1$ degrees of freedom.

Proof. Observe that

- $Z := \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ is a standard normal by Remark 53 and Theorem 55.
- $\chi^2_{n-1} := (n-1)S^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom by Theorem 65.
- Since $\overline{Y} \perp\!\!\!\perp S^2$ by Theorem 65, it follows that $Z \perp\!\!\!\perp \chi^2_{n-1}$ as well.

Hence

$$\begin{aligned}\frac{\overline{Y} - \mu}{S/\sqrt{n}} &= \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\left[(n-1)S^2/\sigma^2\right]/(n-1)}} \\ &:= \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}},\end{aligned}$$

And the right hand satisfies the definition of a t-distribution with $\nu = n-1$ degrees of freedom. □

The third observation in the proof of Theorem 68 uses the following theorem

Theorem 69. *Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ a function only of y. Then the random variables $U = f(X)$ and $V = h(Y)$ are independent.*

Proof idea. We'll show the case where X, Y are discrete random variables, as this illustrate the idea of the general proof. [In lecture, I only sketched this proof, but am including a more detailed version here in the notes for fun.]

We need to show that for all $u, v \in \mathbb{R}$, we have

$$\mathbb{P}\left[U = u, V = v\right] = \mathbb{P}\left[U = u\right]\mathbb{P}\left[V = v\right],$$

as this will establish that U and V are independent.

Let $S_X \subseteq \mathbb{R}$ be the set of values that the random variable X can take, and define S_Y similarly. Let $A_u = \{x \in S_X : g(x) = u\}$ and $B_v = \{y \in S_Y : h(y) = v\}$. Note that the sets A_u and B_v are defined in such a way that

$$[U = u] = [X \in A_u] \quad \text{and} \quad [V = v] = [Y \in A_v] \tag{6}$$

for all $u, v \in \mathbb{R}$. The result then follows from the following calculation:

$$\begin{aligned}
\mathbb{P}[U=u, V=v] &= \mathbb{P}[X \in A_u, Y \in B_v] && \text{by Eq. (6)} \\
&= \mathbb{E}\left[\mathbf{1}_{[X\in A_u, Y\in A_v]}\right] \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} \mathbf{1}_{[X\in A_u, Y\in A_v]}\mathbb{P}[X=x, Y=y] && \text{by Theorem 25} \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} \mathbf{1}_{[X\in A_u, Y\in A_v]}\mathbb{P}[X=x]\,\mathbb{P}[Y=y] && \text{by Theorem 28} \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} \mathbf{1}_{[X\in A_u]}\mathbf{1}_{[Y\in A_v]}\mathbb{P}[X=x]\,\mathbb{P}[Y=y] && \text{since } \mathbf{1}_{E\cap F} = \mathbf{1}_E\mathbf{1}_F \text{ for any events } E, F \\
&= \left(\sum_{x\in S_X} \mathbf{1}_{X\in A_u}\mathbb{P}[X=x]\right)\left(\sum_{y\in S_Y} \mathbf{1}_{Y\in A_u}\mathbb{P}[Y=y]\right) \\
&= \mathbb{P}[X\in A_u]\,\mathbb{P}[Y\in A_v] && \text{by Theorem 21} \\
&= \mathbb{P}[U=u]\,\mathbb{P}[V=v] && \text{by Eq. (6).}
\end{aligned}$$

Since $u, v \in \mathbb{R}$ are arbitrary, this show that U and V are independent. □

Example 70 (Example 7.6 in the textbook)**.** The tensile strengh of a type of wire is normally distributed with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Six pieces of wire are selected from a roll. Let

$$Y_i := \text{the tensile strength for portion } i, \quad i \in [6].$$

We can estimate μ and σ^2 with $\overline{Y}$ and S^2. Because $\sigma^2_{\overline{Y}} = \sigma^2/n$, it follows that $\sigma^2_{\overline{Y}}$ can be estimated by S^2/n.

Question: Find the probability that $\overline{Y}$ is within $2S/\sqrt{n}$ of μ.

$$\begin{aligned}
\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le \frac{2S}{\sqrt{n}}\right] &= \mathbb{P}\left[-\frac{2S}{\sqrt{n}} \le \overline{Y}-\mu \le \frac{2S}{\sqrt{n}}\right] \\
&= \mathbb{P}\left[-2 \le \left(\frac{\overline{Y}-\mu}{S/\sqrt{n}}\right) \le 2\right] \\
&= \mathbb{P}[-2 \le T \le 2],
\end{aligned}$$

where T has a t-distribuion with df $= 5$ (i.e., one less than the number of samples).

A computer tells us that when df $= 5$,

$$\mathbb{P}[-2.015 \le T \le 2.015] = .9$$

Hence, $\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le \frac{2S}{\sqrt{n}}\right]$ is slightly less than .9. In words, the probability that $\overline{Y}$ will be within 2 estimated standard deviations of its mean is slightly less than .90.

Note that if σ^2 were known, the probability that $\overline{Y}$ will fall within $2\sigma_{\overline{Y}}$ of its mean would be given by

$$\begin{aligned}
\mathbb{P}\left[\left|\overline{Y}-\mu\right| \le -2\left(\frac{\sigma}{\sqrt{n}}\right)\right] &= \mathbb{P}\left[-2 \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le 2\right] \\
&= \mathbb{P}[-2 \le Z \le 2] \\
&\approx .95.
\end{aligned}$$

End of Example 70. □

Section 7.2 ends with a discussion about comparing variances of two populations using F statistics. We will skip this, but return to F statistics later.

12 2026-02-06 | Week 04 | Lecture 11

12.1 The central limit theorem

This is based on Section 7.3 and 7.5. The central limit theorem :)

We've seen that if $Y_1, \ldots, Y_n$ is a random sample of normal random variables, then $\overline{Y}$ is normally distributed.

What if the population is not normally distributed?

Main idea of the central limit theorem: If $X_1, \ldots, X_n$ is a random sample from (almost) any distribution, then $\overline{X}$ will be approximately normal when n is large.

Theorem 71 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Let*

$$U_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then

$$U_n \to \mathcal{N}(0,1) \text{ as } n \to \infty,$$

in the sense that

$$\lim_{n\to\infty} \mathbb{P}[U_n \leq u] = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ for all } u \in \mathbb{R}.$$

Remark 72. In practice, this means that when n is large we have $\mathbb{P}[a \leq U_n \leq b] \approx \mathbb{P}[a \leq Z \leq b]$ for any $a, b \in \mathbb{R}$ with $a \leq b$.

The conclusion of the central limit can be summarized in words as the following: $\overline{X}$ is *asymptotically normally distributed with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$*.

Example 73 (Example 7.9 in the textbook)**.** A Ukrainian munitions factory produces a 155mm artillery shell on average every $\mu = 1.5$ minutes, with the production time having a variance of $\sigma^2 = 1$.

Question: Find the probability that the factory will produce more than 100 shells in the next two hours.
Solution: We don't need to assume that the production times of the shells are normal, but we'll assume that they are independent. Let

$$Y_i = (\text{time required to produce the } i^{\text{th}} \text{ shell}).$$

Here $\mu = \mathbb{E}[Y_i] = 1.5$ and $\sigma^2 = \mathrm{Var}(Y_i) = 1$. With this notation, we want to find the probability of the event $[Y_1 + \ldots + Y_{100} \leq 120]$.

$$\begin{aligned}\mathbb{P}[Y_1 + \ldots + Y_{100} \leq 120] &= \mathbb{P}\left[\overline{Y} \leq \frac{120}{100}\right] \\ &= \mathbb{P}\left[\overline{Y} \leq 1.2\right].\end{aligned}$$

Standardizing $\overline{Y}$, we have

$$\begin{aligned}\mathbb{P}\left[\overline{Y} \leq 1.2\right] &= \mathbb{P}\left[\frac{\overline{Y} - 1.5}{1/\sqrt{100}} \leq \frac{1.2 - 1.5}{1/\sqrt{100}}\right] \\ &= \mathbb{P}[U_n \leq 3] \\ &\approx \mathbb{P}[Z \leq -3] && \text{by the CLT} \\ &= 0.0013.\end{aligned}$$

Therefore it is very unlikely that the factory will produce more than 100 shells in the next two hours.

End of Example 73. □

Example 74 (Example 7.10 in the textbook)**.** A candidate believes that she can win a city election if she can earn at least 55% of the votes in precinct 1. She also believes that about 50% of the city's voters favor her. If $n = 100$ voters show up to vote at precinct 1, what is the probability that she will recieve at least 55% of their votes?

Solution: If we think of the $n = 100$ voters as a random sample from the city, we have random variables

$$X_i = \begin{cases} 1 & : \text{ voter i votes for her} \\ 0 & : \text{ otherwise} \end{cases}$$

and the proportion of votes she recieves is

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

We need to find $\mathbb{P}\left[\overline{X} \geq .55\right]$. First observe that

$$\mu := \mathbb{E}\left[X_i\right] = .5 \quad \text{and} \quad \sigma^2 := \text{Var}(X_i) = p(1-p) = (.5)(.5) = .25.$$

By the central limit theorem, since n is large, $\overline{X}$ is approximately normally distributed with mean $\mu_{\overline{X}} = .5$ and variance $\sigma^2_{\overline{X}} = \sigma^2/n = .25/100 = .0025$. Therefore, standardizing $\overline{X}$, we have

$$\begin{aligned} \mathbb{P}\left[\overline{X} \geq .55\right] &= \mathbb{P}\left[\frac{\overline{X} - .5}{\sqrt{.0025}} \geq \frac{.55 - .5}{\sqrt{.0025}}\right] \\ &= \mathbb{P}\left[U_n \geq 1\right] \\ &\approx \mathbb{P}\left[Z \geq 1\right] && \text{by the CLT} \\ &= 0.1587. \end{aligned}$$

End of Example 74. □

12.2 Point Estimation

This section is based on section 8.2 in the textbook.

Suppose we wish to estimate a target parameter θ, say, the fraction of people living in Hawai'i who like broccoli.

Remark 75. If our target parameter is θ, we often denote our estimator for that parameter using a "hat", i.e., as $\widehat{\theta}$.

Recall that the distribution of an estimator $\widehat{\theta}$ is called the **sampling distribution**. It is desirable for the sampling distribution to cluster around the target parameter.

One criterion of the "goodness" of an estimator is to ask whether it exhibits "bias".

Definition 76 (Biased and unbiased)**.** Let $\widehat{\theta}$ be a point estimator for a parameter $\theta \in \mathbb{R}$.

If $\mathbb{E}\left[\widehat{\theta}\right] = \theta$ then $\widehat{\theta}$ is ***unbiased.***

If $\mathbb{E}\left[\widehat{\theta}\right] \neq \theta$, then $\widehat{\theta}$ is said to be ***biased***.

The ***bias*** of an estimator $\widehat{\theta}$, denoted $B(\widehat{\theta})$ or $\text{Bias}(\widehat{\theta})$, is the number $B(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$.

Remark 77. Since $\widehat{\theta}$ is a statistic, it is a function of a random sample. So when we write $\mathbb{E}[\widehat{\theta}]$, we really mean $\mathbb{E}\left[\widehat{\theta}(X_1, \ldots, X_n)\right]$, where $X_1, \ldots, X_n$ is a random sample from the population.

Remark 78. Confusingly, when given an estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ (a random variable which depends on the sample), we also sometimes denote the value of resulting estimate by $\widehat{\theta}$.

Example 79 (Estimating the sides of a dice)**.** Suppose I roll a dice behind a screen, shouting out the number rolled each time. Your job is to estimate the number of sides of the dice, which we will call θ.

Given n rolls, the sample consists of the n random variables

$$X_1, \ldots, X_n$$

where X_i is the i^{th} dice roll. We'll give two examples of estimators:

- Our first example is called the *running maximum*, and it is defined as

  $$\widehat{\theta} = \max(X_1, \ldots, X_n).$$

  For example, if I roll the dice four times and obtain $3, 11, 7$, and 4, then your estimate would be that the dice has 11 sides.

  This sounds silly, but with enough data this estimator will give you the correct answer with high probability. This is because, if I roll the dice enough times, then I will eventually roll the highest number; hence,

  $$\lim_{n\to\infty} \widehat{\theta}(X_1, \ldots, X_n) = \theta.$$

  One drawback is that this estimator is biased because $\mathbb{E}[\widehat{\theta}] < \theta$. This is because the estimator will "usually" be less than the true number of sides, and will never be more—so on average, $\widehat{\theta}$ will be smaller than θ.

- Our second example is what is called the *method of moments estimator*. In this setting, it's defined as

  $$\widehat{\theta} = 2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1.$$

  Let's demystify this formula. Suppose X a random variable representing one roll of the dice. We know that since the dice has θ sides, X has expected value

  $$\mathbb{E}[X] = \frac{1}{\theta}(1 + 2 + \ldots + \theta) = \frac{\theta + 1}{2}.$$

  Rearranging, we get

  $$\theta = 2\mathbb{E}[X] - 1. \tag{7}$$

  We also know that by the law of large numbers, $\mathbb{E}[X] \approx \frac{X_1,\ldots,X_n}{n}$. Combining this with Eq. (7) implies

  $$\theta \approx 2\left(\frac{X_1, \ldots, X_n}{n}\right) - 1.$$

  The right hand side is what we've chosen to be our $\widehat{\theta}$, and hopefuly this calculation makes it seem like a reasonable choice.

  This estimator is unbiased. To see why, observe that

  $$\begin{aligned}
  \mathbb{E}[\widehat{\theta}] &= \mathbb{E}\left[2\left(\frac{X_1 + \ldots + X_n}{n}\right) - 1\right] \\
  &= 2\mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] - 1 \\
  &= 2\mathbb{E}[X] - 1 \\
  &= \theta && \text{by Eq. (7).}
  \end{aligned}$$

  In our example, with dice rolls $3, 11, 7$, and 4, our estimator $\widehat{\theta}$ returns the estimate

  $$2\left(\frac{3 + 11 + 7 + 4}{4}\right) - 1 = 2\left(\frac{25}{4}\right) - 1 = 11.5.$$

  So your estimate would be that the dice has 11.5 sides.

End of Example 79. □

Given two unbiased estimators, then *ceteris paribus*, we would typically prefer the estimator which has lower variance.

13 2026-02-11 | Week 05 | Lecture 12

Please read 8.2, 8.3, and 8.4.

13.1 Mean square error

This section is based on chapter 8.2 and 8.4.

Here we present another criterion for whether an estimator is "good".

Definition 80 (Mean square error)**.** The ***error of estimation*** ϵ is the distance between an estimator and its target parameter; that is, $\epsilon = |\widehat{\theta} - \theta|$. The ***mean square error*** (MSE) of a point estimator $\widehat{\theta}$ is

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[|\widehat{\theta} - \theta|^2\right]$$

The error of estimation is a random variable, so we cannot say how large or small it will be for a particular estimate. But on average, we prefer estimators which are, in theory, close to the target parameter on average. This is what the what the MSE is measuring: it is *the average square of the distance between the estimator and the target parameter*. We want the MSE to be small.

The MSE is a useful measure because, as the next theorem shows, it is a combined measure of both the estimator's variance and its bias (both of which we want to be small).

Theorem 81. *Let $\widehat{\theta}$ be an estimator such that $\mathbb{E}[\widehat{\theta}]$ is defined. Then,*

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \left[\mathrm{Bias}(\widehat{\theta})\right]^2.$$

Proof. Let $B = \mathrm{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$. Note that B is nonrandom; it is just some number. Write

$$\widehat{\theta} - \theta = \widehat{\theta} - \mathbb{E}[\widehat{\theta}] + \mathbb{E}[\widehat{\theta}] - \theta = \left[\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right] + B.$$

Hence

$$\left(\widehat{\theta} - \theta\right)^2 = \left(\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right)^2 + B^2 + 2\left(\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right)B.$$

Taking expectations of both sides gives

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right)^2\right] + \mathbb{E}\left[B^2\right] + 2\mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right)B\right] \\ &= \mathrm{Var}(\widehat{\theta}) + B^2 + 2B\underbrace{\mathbb{E}\left[\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right]}_{=0}\end{aligned}$$

where the last step follows by linearity of expectation: $\mathbb{E}\left[\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right] = \mathbb{E}[\widehat{\theta}] - \mathbb{E}[\widehat{\theta}]$. □

13.2 Standard Error

Based on sections 8.3 and 8.4 in the textbook

Definition 82 (Standard Error)**.** Given a point estimator $\widehat{\theta}$, the ***standard error*** $\sigma_{\widehat{\theta}}$ is the standard deviation of its sampling distribution; that is,

$$\sigma_{\widehat{\theta}} = \sqrt{\mathrm{Var}(\widehat{\theta})}.$$

We've already seen one example of standard error (in the denominator of U_n in the central limit theorem), which we state plainly in the following proposition:

Proposition 83 (Standard error of a sample mean)**.** *Let $X_1, \ldots, X_n$ be a random sample from a population with variance σ^2. Then*

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Proof. Using Theorem 33,

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) \\
&= \frac{1}{n^2}\left(n\sigma^2\right) \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Taking square roots implies the statement of the proposition. □

The following is an example of a *concentration inequality*, i.e., a probabilisitic inequality which controls the deviation of a random variable from its mean.

Theorem 84 (Chebychev's Inequality)**.** *Let Y be a random variable with mean μ and variance σ^2. Then for any $k > 0$,*

$$\mathbb{P}\left[|Y - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}.$$

Proof. The proof follows from an application of Markov's inequality (Theorem 35), which says that for any nonnegative random variable X and scalar $a > 0$,

$$\mathbb{P}\left[X \geq a\right] \leq \frac{\mathbb{E}\left[X\right]}{a}. \tag{8}$$

To see this, observe that

$$\begin{aligned}
\mathbb{P}\left[|Y - \mu| \geq k\sigma\right] &= \mathbb{P}\left[|Y - \mu|^2 \geq (k\sigma)^2\right] \\
&\leq \frac{\mathbb{E}\left[|Y - \mu|^2\right]}{k^2\sigma^2} && \text{by Eq. (8) with } X = Y - \mu \text{ and } a = (k\sigma)^2 \\
&= \frac{\sigma^2}{k^2\sigma^2} && \text{by definition of variance} \\
&= \frac{1}{k^2}.
\end{aligned}$$

□

Remark 85. The trick of squaring both sides of the inquality and then applying Markov's inequality in the proof of Theorem 84 was also used in the proof of the weak law of large numbers (Theorem 36).

Since $\widehat{\theta}$ is a random variable, the error of estimation $\epsilon = |\widehat{\theta} - \theta|$ is also a random quantity, so we cannot say how large or small it will be for a particular estimate. But we can make probabilistic statements about it. Here is one such example.

Proposition 86. *Suppose $\widehat{\theta}$ is an unbiased estimator of θ with finite variance, and let $\epsilon = |\widehat{\theta} - \theta|$ be the estimation error. Then*

$$\mathbb{P}\left[\epsilon \leq k\sigma_{\widehat{\theta}}\right] \geq 1 - \frac{1}{k^2}.$$

Proof.

$$\begin{aligned}
\mathbb{P}\left[\epsilon \leq k\sigma_{\widehat{\theta}}\right] &= \mathbb{P}\left[|\widehat{\theta} - \theta| \leq k\sigma_{\widehat{\theta}}\right] \\
&= \mathbb{P}\left[\left|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right| \leq k\sigma_{\widehat{\theta}}\right] && \text{since } \theta = \mathbb{E}[\widehat{\theta}] \\
&= 1 - \mathbb{P}\left[\left|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\right| \geq k\sigma_{\widehat{\theta}}\right] \\
&\geq 1 - \frac{1}{k^2} && \text{by Theorem 84.}
\end{aligned}$$

□

The textbook suggests taking $k = 2$, which gives

$$\mathbb{P}\left[\text{estimation error} \leq 2\sigma_{\widehat{\theta}}\right] \geq .75. \tag{9}$$

so the error of estimation ϵ will be less than $2\sigma_{\widehat{\theta}}$ with probability at least .75. This bound is conservative; for many random variables the true probability will be much higher. For example, it will be around .95 when $\widehat{\theta}$ is approximately normally distributed (e.g., when $\widehat{\theta}$ is a sample mean).

14 2026-02-13 | Week 05 | Lecture 13

14.1 k-standard-error bounds

This is based on section 8.4 of the textbook.

Definition 87 (k-standard-error bound)**.** Let $\widehat{\theta}$ be an point estimator of a target parameter $\theta \in \mathbb{R}$, and let $k > 0$. A ***k*-standard-error bound** for the estimator $\widehat{\theta}$ is an inequality of the form

$$\mathbb{P}\left[|\widehat{\theta} - \theta| \leq 2\sigma_{\widehat{\theta}}\right] \geq 1 - \alpha \tag{10}$$

where $\alpha \in [0, 1]$.

Remark 88. Recall that $\epsilon = |\widehat{\theta} - \theta|$ is the *error of estimation*, so the left hand side of Eq. (10) can be understood as "the probability that the estimation error is small". As such, when establishing a k-standard-error bound, we want $1 - \alpha$ to be as close to 1 as possible.

In the last lecture, we used Chebychev's inequality to establish the following 2-standard-error bound for any unbiased estimator $\widehat{\theta}$ that has finite variance:

$$\mathbb{P}\left[|\widehat{\theta} - \theta| \leq 2\sigma_{\widehat{\theta}}\right] \geq .75. \tag{11}$$

This inequality says that the error of estimation will be less than $2\sigma_{\widehat{\theta}}$ with probability at least $1 - \alpha = .75$. This bound is conservative, since $1 - \alpha = .75$ is not very close to 1. Usually the true value of $1 - \alpha$ is much higher. For example, $1 - \alpha$ will be around .95 when $\widehat{\theta}$ is approximately normally distributed (e.g., when $\widehat{\theta}$ is a sample mean). We show that with the next example.

Example 89 (Example 8.2 in textbook)**.** A sample of $n = 1000$ randomly selected voters showed $y = 560$ in favor of candiate Jones. Estimate p, the faction of voters in the population favoring Jones, and place a 2-standard-error bound on the error of estimation (i.e., construct an inequality of the form Eq. (10)).

Solution: Let $\widehat{p}$ be the fraction of of the sample who support Jones. According to this estimator, our estimate is that 56% of voters support Jones. How much faith should we put in this value?

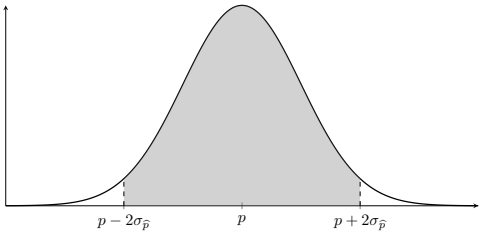
To begin with, observe that

$$\widehat{p} = \frac{X_1 + \ldots + X_{1000}}{n}$$

where

$$X_i = \begin{cases} 1 & : \text{ voter i supports Jones} \\ 0 & : \text{ otherwise} \end{cases}$$

Since $\widehat{p}$ is a sample average with $n = 1000$, the central limit theorem says that its distribution is approximately Gaussian with mean p (since sample averages are unbiased). Thus, the sampling distribution of $\widehat{p}$ looks something like this:



The sampling distribution of $\widehat{p}$

The question is asking us to do two things: (1) find $\sigma_{\widehat{p}}$ and (2) find the area of the grey shaded region. To do this, let's first compute the standard error of $\sigma_{\widehat{p}}$.

Observe that $\sigma^2 := \mathrm{Var}(X_i) = p(1-p)$, so $\sigma = \sqrt{p(1-p)}$ *(This was HW 3 exercise 1a).* Therefore by Proposition 83,

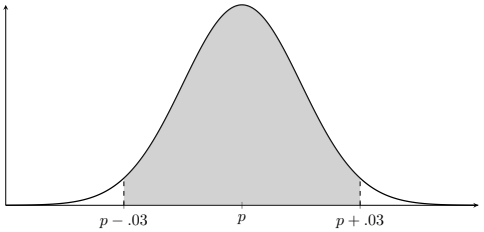
$$\sigma_{\widehat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{1000}}.$$

Oops! We don't know p, since that's what we're trying to estimate. Instead, using the approximation $p(1-p) \approx \widehat{p}(1-\widehat{p})$, we approximate $\sigma_{\widehat{p}}$ as the following:

$$\sigma_{\widehat{p}} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{1000}} = \sqrt{\frac{(.56)(.44)}{1000}} = .015. \tag{12}$$

(There are other things we could've done here, but I'm following the approach in the textbook for this example.)

We can now refine our picture a bit:



The sampling distribution of $\widehat{p}$

It remains to estimate the area of the shaded region, which represents the probability that the estimation error is less than .03. As a first approximation, we can apply the Chebychev bound of Eq. (9), which gives

$$\mathbb{P}\left[\text{estimation error} \leq .03\right] \approx \mathbb{P}\left[|\widehat{p} - p| \leq 2\sigma_{\widehat{p}}\right] \geq .75.$$

Hence, the shaded region has area at least .75. But this is very conservative.

Since we know that $\widehat{p}$ has an approximately Gaussian sampling distribution, we can use that information to do can do better than .75, as follows:

First, since $\widehat{p}$ is a sample mean of $X_1, \ldots, X_{1000}$, the central limit theorem (Theorem 71) implies that

$$\frac{\widehat{p} - p}{\sigma_{\widehat{p}}} \approx Z, \tag{13}$$

where Z has a standard normal distribution. Hence,

$$\begin{aligned}
\mathbb{P}\left[\text{estimation error} \leq .03\right] &= \mathbb{P}\left[|\widehat{p} - p| \leq 2\sigma_{\widehat{p}}\right] && \text{since } \sigma_{\widehat{p}} \approx .015 \text{ by Eq. (12)} \\
&= \mathbb{P}\left[-2\sigma_{\widehat{p}} \leq \widehat{p} - p \leq 2\sigma_{\widehat{p}}\right] \\
&= \mathbb{P}\left[-2 \leq \frac{\widehat{p} - p}{\sigma_{\widehat{p}}} \leq 2\right] && \text{standardizing} \\
&\approx \mathbb{P}\left[-2 \leq Z \leq 2\right] && \text{by Eq. (13)} \\
&= .95.
\end{aligned}$$

Thus, if we were to repeat this poll 100 times, we would expect that only in about 5 such polls would the estimate deviate from the true p by more than .03. So we can be reasonably confident that our estimate of $\widehat{p} = .56$ is within .03 of the true value of p.

End of Example 89. □

15 2026-02-23 | Week 07 | Lecture 14

15.1 Confidence intervals

This section is based on chapter 8.5 in the textbook.

Definition 90 (Interval estimator)**.** An ***interval estimator*** is a rule specifying a way to calculate two statistics that form endpoints of an interval. Interval estimators are also called ***confidence intervals***.

In other words, a confidence interval is an interval $[\widehat{\theta}_L, \widehat{\theta}_U]$, where $\widehat{\theta}_L$ and $\widehat{\theta}_U$ are a pair of random variables (which vary from sample to sample).

Remark 91. Desired properties:

- the interval contains the target parameter θ
- the interval is relatively narrow

The upper and lower endpoints of a confidence interval are called the ***upper*** and ***lower confidence limits***, respectively. The probability that a (random) confidence interval will enclose the target parameter θ is called the ***confidence coefficient***. If this is high, then we can be highly confident that a confidence interval constructed using the results of a sample, will enclose θ.

Put differently, suppose $\widehat{\theta}_L$ and $\widehat{\theta}_U$ are the upper and lower confidence limits for a parameter θ. If

$$\mathbb{P}\left[\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U\right] = 1 - \alpha,$$

then $(1-\alpha)$ is the ***confidence coefficient***.

Example 92 (Example 8.4 in textbook)**.** Suppose we obtain a single observation Y from an exponential distribution with mean $\theta > 0$. Use Y to form a confidence interval for θ with confidence coefficient 0.90.

Solution: First, recall that a random variable X is an exponetial with mean $\mu > 0$ if and only if

$$\mathbb{P}[X > t] = e^{-\frac{1}{\mu}t} \tag{14}$$

for every $t > 0$.

Claim 1: $U = Y/\theta$ is a mean 1 exponential random variable

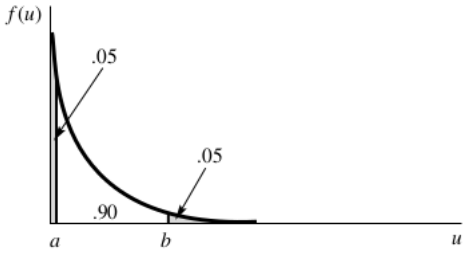
Proof of Claim 1. Let $t > 0$. Then

$$\begin{aligned}
\mathbb{P}[U > t] &= \mathbb{P}\left[\frac{Y}{\theta} > t\right] \\
&= \mathbb{P}[Y > \theta t] \\
&= e^{-\frac{1}{\theta}(\theta t)} && \text{by Eq. (14)} \\
&= e^{-t}
\end{aligned}$$

By Eq. (14) we conclude that U is a mean 1 exponential. □ Claim

Remark 93. In class, I claimed that we also need to consider the case $t \leq 0$ in the proof of Claim 1. This was incorrect. It is not necessary to do so because Eq. (14) is an "if and only if" statement.

As a mean 1 exponential, U has density $f(u) = e^{-u}\mathbf{1}_{[u>0]}$. Critically, the distribution of U does not depend on the unknown parameter θ, and we can find real numbers $a, b > 0$ such that



So that

$$\mathbb{P}\left[a \leq U \leq b\right] = \int_a^b e^{-u} du = .90.$$

The numbers are $a = .051$ and $b = 2.995$. Hence

$$\begin{aligned} .90 &= \mathbb{P}\left[.051 \leq \frac{Y}{\theta} \leq 2.996\right] \\ &= \mathbb{P}\left[\frac{Y}{2.996} \leq \theta \leq \frac{Y}{.051}\right] \qquad \text{using algebra and } Y \geq 0. \end{aligned}$$

Thus, our 90% confidence interval estimator is

$$[\widehat{\theta}_L, \widehat{\theta}_U] = \left[\frac{Y}{2.996}, \frac{Y}{.051}\right].$$

Story: Suppose you decide to spend your Saturday standing at an idyllic country road waiting for cars to pass, measuring the time between cars. We assume the inter-arrival times between cars is exponentially distributed. The unknown parameter θ is the average time between cars in, say, minutes. Suppose the time between the first two cars to pass is $Y = 12$ minutes. Then we compute

$$\widehat{\theta}_L = \frac{12}{2.996} \approx 4 \quad \text{and} \quad \widehat{\theta}_U = \frac{9}{.051} \approx 240.$$

The conclusion is that with probability 90%, the next car will take between 4 minutes and 4 hours to arrive. This range of this interval is so wide because it's based on very little data—only one sample.

End of Example 93. □

Remark 94 (Pivotal method)**.** Example 92 illustrate the "pivotal method" of computing confidence intervals, which has two steps:

1. Construct a function of the sample and the target parameter θ whose distribution does not depend on θ. This is called the "pivotal quantity". In Example 92 the pivotal quantity was $U = Y/\theta$.
2. Use known facts about the distribution of U to derive a confidence interval. In Example 92, we used the fact that
$$\mathbb{P}\left[.051 \leq U \leq 2.996\right] = .90.$$

We've discussed ***2-sided confidence intervals***, which take the form $[\widehat{\theta}_L, \widehat{\theta}_U]$. A ***lower 1-sided confidence interval*** takes the form $[\widehat{\theta}_L, \infty)$, and an ***upper 1-sided confidence*** interval takes the form $(-\infty, \widehat{\theta}_U]$. The theory for 1-sided confidence intervals is similar to that of 2-sided confidence intervals.

15.2 Large-sample confidence intervals

This section is based on chapter 8.6 in the textbook.

Suppose we wish to construct an $100(1-\alpha)\%$ confidence interval for a target parameter θ, where $\alpha \in (0, 1)$. For doing this, there are two settings which are often qualitatively different:

- When the sample size n is small
- When the sample size n is large

We'll treat the large n case first. In this case, it is commonly the case that one can find a point estimator $\widehat{\theta}$ which is approximately normal (usually due to the central limit theorem), in which case the following proposition can be used to construct a confidence interval.

Proposition 95 (Large n confidence intervals – This is example 8.6 in the textbook)**.** *Suppose that $\widehat{\theta}$ is an unbiased estimator of θ. Assume that $\widehat{\theta}$ is approximately normal (a common situation when the sample size n is large), and denote the standard error of $\widehat{\theta}$ by $\sigma_{\widehat{\theta}}$. Let $\alpha \in (0,1)$ be arbitrary, and choose a number $z_{\alpha/2} \in \mathbb{R}$ such that*

$$\mathbb{P}\left[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right] = 1 - \alpha. \tag{15}$$

Then the interval

$$[\widehat{\theta}_L, \widehat{\theta}_U] := [\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}},\ \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}]$$

forms a $100(1-\alpha)\%$ confidence interval for θ.

Proof. Since $\widehat{\theta}$ is unbiased, $\mathbb{E}[\widehat{\theta}] = \theta$. Therefore, by the central limit theorem / standardization,

$$\frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \approx Z. \tag{16}$$

Plugging Eq. (16) into Eq. (15),

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \leq z_{\alpha/2}\right] \approx 1 - \alpha.$$

Rearranging terms in the inequalities gives

$$\mathbb{P}\left[\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} \leq \theta \leq \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right] \approx 1 - \alpha,$$

which implies that statement of the proposition. □

16 2026-02-25 | Week 07 | Lecture 15

Section 8.6 in text book

16.1 Large sample confidence intervals (continued)

16.1.1 Comparing two populations

Example 96 (Example 8.8 in textbook)**.** Two brands of refridgerators, denoted B_1 and B_2, are tested. In a random sample of $n_1 = 50$ fridges from brand B_1, 12 failed within the first year. In a random sample of $n_2 = 60$ fridges from brand B_2, 12 failed within the first year.

Let p_i be the proportion of brand B_i fridges that fail in the first year, and let $\widehat{p}_i$ be the observed proportion of the sampe which do so. Estimate the difference $p_1 - p_2$, with confidence coefficient approximately .98.

Solution: Let $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$. We will apply Proposition 95 with confidence coefficient $1 - \alpha = .98$, so $\alpha = .02$. We need

- the value of the estimate $\widehat{\theta}$
- the value of the standard error $\sigma_{\widehat{\theta}}$
- the value of $z_{0.01}$

First, the estimate is

$$\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2 = \frac{12}{50} - \frac{12}{60} = .04.$$

Second, the standard error is

$$\begin{aligned}
\sigma_{\widehat{\theta}} &= \sqrt{\mathrm{Var}(\widehat{\theta})} \\
&= \sqrt{\mathrm{Var}(p_1) + \mathrm{Var}(p_2)} \\
&= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
&= \sqrt{\frac{p_1(1-p_1)}{50} + \frac{p_2(1-p_2)}{60}}.
\end{aligned}$$

Oops! We don't know p_1 or p_2. Luckily, we can use the approximation $p_i(1-p_i) \approx \widehat{p}_i(1-\widehat{p}_i)$ for $i = 1, 2$, we have

$$\begin{aligned}
\sigma_{\widehat{\theta}} &\approx \sqrt{\frac{(.24)(.76)}{50} + \frac{(.20)(.80)}{60}} \\
&= .0791
\end{aligned}$$

Finally, $z_{0.01} = 2.33$ (draw a picture).

Putting these three things together and applying Proposition 95,

$$[\widehat{\theta}_L, \widehat{\theta}_U] = [.04 - (2.33)(.0791),\ .04 + (2.33)(.0791)] = [-.144,\ .224]$$

is a confidence interval for the difference in proportions $p_1 - p_2$. Since this confidence interval contains zero, it is believable that $p_1 = p_2$. But it also contains 1, so a value of $p_1 - p_2 = .1$ is also a believable value.

To be conservative in our judgement, we would say that we cannot reject the possibility that the brands perform equally well.

End of Example 96. □

16.1.2 Sample complexity

This part is based on Section 8.7

One of the most common questions that researchers face when designing experiments is "how many samples do I need?". This is the topic of *sample complexity*, which is a rich and beautiful subject in statistics, but can be hard to wrap your head around at first. Here's an example.

Example 97 (This is like Example 8.9 in the textbook)**.** A pollster is hired to estimate the proportion p of people who support candidate Jones in the upcoming election. The pollster wants to have 90% confidence that their poll is no greater than 4 percentage points away from the true value of p. How many people must they poll?

Solution: We'll take as our estimator the sample mean $\widehat{p}$:

$$\widehat{p} = \frac{\xi_1 + \ldots + \xi_n}{n}, \quad \text{where} \quad \xi_i = \begin{cases} 1 & : \text{ person } i \text{ supports candidate Jones} \\ 0 & : \text{ otherwise} \end{cases},$$

and $n > 1$ is to be determined later. By the Law of Large Numbers, $\widehat{p} \to p$ as n grows, which implies

$$\mathbb{P}\left[|\widehat{p} - p| \leq .04\right] \to 1 \quad \text{as } n \to \infty.$$

Hence we need to find n sufficiently large that

$$\mathbb{P}\left[|\widehat{p} - p| \leq .04\right] \geq 0.90.$$

In other words, our goal is to find a value for n such that the interval

$$[\widehat{p} - .04,\ \widehat{p} + .04] \tag{17}$$

is a 90% confidence interval for p.

We'll assume that the poll is sufficiently large that $\widehat{p}$ is approximately normally distributed (a good rule of thumb for "sufficiently large" is that $\min\{np, n(1-p)\} \geq 10$; usually $n = 30$ is sufficient for the approximation to be good enough. Here we'll definitely be polling at least 30 people, which is enough to conclude that $\widehat{p}$ is approximately normal). Make two observations:

- First, observe that
$$\mathbb{P}\left[-1.6 \leq Z \leq 1.6\right] \approx 0.90.$$
Therefore, since $\widehat{p}$ is approximately normal, Proposition 95 implies that the interval
$$[\widehat{p} - 1.6\sigma_{\widehat{p}},\ \widehat{p} + 1.6\sigma_{\widehat{p}}]$$
is a 90% confidence interval for p.
- Second, observe that using $\mathrm{Var}(\xi_i) = p(1-p)$, we can show that
$$\sigma_{\widehat{p}} = \sqrt{\mathrm{Var}(\widehat{p})} = \sqrt{\mathrm{Var}\left(\frac{\xi_1 + \ldots + \xi_n}{n}\right)} = \sqrt{\frac{p(1-p)}{n}}.$$

Therefore the interval

$$\left[\widehat{p} - 1.6\sqrt{\frac{p(1-p)}{n}},\ \widehat{p} + 1.6\sqrt{\frac{p(1-p)}{n}}\right]$$

is a 90% confidence interval for p. In order for this to take the desired form $[\widehat{p} - .04,\ \widehat{p} + .04]$ or narrower, we need to find the value of n such that

$$1.6\sqrt{\frac{p(1-p)}{n}} \leq .04.$$

Rearranging terms gives

$$n \geq 1600p(1-p). \tag{18}$$

Oops! We don't know p. But there's a fix. Observe that $1600p(1-p) \leq 1600 \cdot \frac{1}{4} = 400$ for any $p \in (0,1)$. Therefore, for any value of p, a sample size of at least $n = 400$ voters is sufficient (since $n \geq 400$ implies Eq. (18) holds). Therefore the pollster needs to poll at least 400 people in order to have 90% confidence that their estimate $\widehat{p}$ is within 4 percentage points of the true proportion of people who support candidate Jones.

(Using the more accurate value of 1.645 rather than 1.6 gives the requirement of $n \geq 423$.)

End of Example 97. □

17 2026-02-27 | Week 07 | Lecture 16

This lecture is based on Chapter 8.8 in the textbook.

17.1 Small-sample confidence intervals

Recall from Theorem 68 that if $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

follows a Student t-distribution with $n-1$ degrees of freedom.

Example 98 (Similar to example 8.11 in the textbook)**.** The muzzle velocities of a sample of eight high-velocity artillery shells are observed to be

972	960	965	982
1006	1053	1056	1062

(measured in meters per second). The sample mean is $\overline{Y} = 1007$ meters per second.

Question: Find a 95% confidence interval for the true average velocity μ of shells of this type. Assume muzzle velocities are normally distributed.

Solution: Our samples are $Y_1, \ldots, Y_8$. Since $n = 8$, so the large-sample techniques discussed earlier are inapplicable. Let

$$\overline{Y} = \frac{Y_1 + \ldots + Y_8}{8}.$$

Then $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/8)$ since $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

If we were to standardize $\overline{Y}$, we'd have:

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{8}} = Z \sim \mathcal{N}(0, 1)$$

But since we don't know σ, we can't actually standardize. Instead, we do the "approximate standarization" technique discussed in Section 11.1 (and motivated in Section 10.1), whereby we divide by the sample standard deviation S rather than the unknown population standard deviation σ:

$$\overline{Y} \mapsto \frac{\overline{Y} - \mu}{S/\sqrt{8}}$$

This transformation—when we divided by S rather than σ—is called ***studentization*** (as opposed to ***standardization*** which involves dividing by σ). Recalling Theorem 68, it follows that the "studentized" statistic

$$T := \frac{\overline{Y} - \mu}{S/\sqrt{8}} \tag{19}$$

has a t-distribution with $n - 1 = 7$ degrees of freedom.

We can now use the pivotal method (described in Remark 94) to construct a 95% confidence interval for μ; in this instance our pivotal quantity is T, as its distribution does not depend on the parameter μ.

Using the `R` quantile function `qt(.025,df=7,lower.tail=TRUE)` and the symmetry of the t-distribution, we see that

$$\mathbb{P}\left[-2.365 \leq T \leq 2.365\right] = 0.95$$

Therefore, we have:

$$\begin{aligned} 0.95 &= \mathbb{P}\left[-2.365 \leq T \leq 2.365\right] \\ &= \mathbb{P}\left[-2.365 \leq \frac{\overline{Y}-\mu}{S/\sqrt{8}} \leq 2.365\right] && \text{by Eq. (19)} \\ &= \mathbb{P}\left[\overline{Y} - \frac{2.365S}{\sqrt{8}} \leq \mu \leq \overline{Y} + \frac{2.365S}{\sqrt{8}}\right] && \text{by doing algebra} \end{aligned}$$

Therefore, the interval $[\overline{Y} - \frac{2.365S}{\sqrt{8}},\ \overline{Y} + \frac{2.365S}{\sqrt{8}}]$ is a 95% confidence interval for μ.

For our data, we have $\overline{Y} = 1007$, and by direct computation one can show that $S = 43.7$. Plugging these values in gives the observed confidence interval

$$[970.5, 1043.5]$$

for our sample.

End of Example 98. □

Remark 99. If we had (falsely) assumed that T from Eq. (19) had a standard normal distribution, then our confidence interval would have been

$$\left[\overline{Y} - \frac{1.96S}{\sqrt{8}},\ \overline{Y} + \frac{1.96S}{\sqrt{8}}\right] = [977, 1037].$$

This is narrower (which may seem good) but this is because we wouldn't be accounting for the higher variance that $\overline{Y}$ has due our small sample size. As the result, the interval wouldn't truly be a 95% interval. It might actually be an 80% interval or something—so our confidence in our estimate would be misplaced.

17.2 Statistical consistency

(This section is based on Chapter 9.3 in the textbook.)

Our next big topic after confidence intervals will be properties of estimators. There are three big ones:

- Consistency (a benchmark that an estimator should satisfy)
- Relative efficiency (a criteron for comparing estimators)
- Sufficiency (an information-theoretic criterion for statistics in general to consider when constructing estimators)

Definition 100 (Convergence in probability)**.** Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables and let $c \in \mathbb{R}$. We say that the sequence (X_n) ***converges in probability*** to c if

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - c| > \epsilon\right] = 0 \tag{20}$$

for every $\epsilon > 0$. We denote convergence in probability by writing $X_n \xrightarrow{P} c$.

More generally, if X is a random variable, we say that (X_n) ***converges in probability*** to X if

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \epsilon\right] = 0$$

for every $\epsilon > 0$. In this case we write $X_n \xrightarrow{P} X$.

Remark 101 (Interpretation of Definition 100)**.** An equivalent formulation of Eq. (20) is

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - c| \leq \epsilon\right] = 1.$$

In other words, $X_n \xrightarrow{P} c$ means that for large n, X_n will be close to c with high probability.

Definition 102 (Statistical consistency)**.** Let θ be a parameter and let $\widehat{\theta}_n$ be an estimator of θ which is calculated using a sample of size n. We say that $\widehat{\theta}_n$ is ***consistent*** if $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

You've already seen an important example of a consistent estimator: the sample mean.

Example 103 (Sample mean is consistent)**.** If $X_1, \ldots, X_n$ is a random sample from a distribution with finite variance and expectation μ, then $\overline{X} = \frac{X_1 + \ldots + X_n}{n}$ is a consistent estimator of μ by the weak law of large numbers (Theorem 36).

End of Example 103. □

Another example I won't discuss in detail is the following.

Example 104. If T_n is a t-distribution with n degrees of freedom, and $Z \sim \mathcal{N}(0,1)$ then

$$T_n \xrightarrow{P} Z \text{ as } n \to \infty.$$

I mentioned something about this in Section 10.3.

End of Example 104. □

Here is another example

Example 105 (Consistency of the running maximum)**.** Fix $\theta > 0$, and let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution on the interval $(0, \theta)$. Suppose we wish to estimate θ. One natural estimator is the running maximum, defined as $M = \max(Y_1, \ldots, Y_n)$. We will show that M is a *consistent* estimator of θ.

Let F denote the cumulative distribution function of M; i.e., $F(t) := \mathbb{P}[M \leq t]$ for all $t \in \mathbb{R}$.

Claim 1: $F(t) = \left(\frac{t}{\theta}\right)^n \mathbf{1}_{[t \in (0,\theta)]}$ for all $t \in \mathbb{R}$.

Proof of Claim 1. There are three cases, depending on the value of t:

- *Case 1.* If $t \leq 0$ then $F(t) = 0$. (This is obvious.)
- *Case 2.* If $t \geq 1$ then $F(t) = 1$. (This is also obvious.)
- *Case 3.* If $t \in (0, \theta)$ then

$$\begin{aligned}
F(t) &= \mathbb{P}[M \leq t] && \text{by definition of cdf} \\
&= \mathbb{P}[\max(Y_1, \ldots, Y_n) \leq t] && \text{by definition of } M \\
&= \mathbb{P}[Y_1 \leq t, \ldots, Y_n \leq t] && \text{think about it} \\
&= \prod_{i=1}^{n} \mathbb{P}[Y_i \leq t] && \text{by independence of } Y_1, \ldots, Y_n \\
&= \left(\frac{t}{\theta}\right)^n && \text{since } \mathbb{P}[Y_i \leq t] = \frac{t}{\theta} \text{ when } t \in (0, \theta)\ .
\end{aligned}$$

Putting these three cases together implies the statement of the claim. □ Claim

Claim 2: $M \xrightarrow{P} \theta$ as $n \to \infty$.

Proof of Claim 2. We need to show that $\mathbb{P}[|M - \theta| > \epsilon] \to 0$ as $n \to \infty$ for every $\epsilon > 0$.

Let $\epsilon > 0$ be arbitrary. Without loss of generality, we may assume that $\epsilon < \theta$. Then

$$\begin{aligned}
\mathbb{P}[|M - \theta| > \epsilon] &= \mathbb{P}[M < \theta - \epsilon] && \text{since } M < \theta \text{ with probability 1} \\
&= F(\theta - \epsilon) \\
&= \left(\frac{\theta - \epsilon}{\theta}\right)^n .
\end{aligned}$$

Since $0 < \frac{\theta - \epsilon}{\theta} < 1$, the right-hand side converges to zero as $n \to \infty$. □ Claim

By Claim 2, M is a consistent estimator of θ.

End of Example 105. □

18 2026-03-02 | Week 08 | Lecture 17

18.1 Relative efficiency

Based on section 9.2 in the textbook.

Recall that if we have two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of the same target parameter θ, we prefer the one with the smaller variance.

Definition 106 (Relative efficiency)**.** If $\widehat{\theta}_1, \widehat{\theta}_2$ are unbiased estimators of a parameter θ, then the ***efficiency*** of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$, is

$$\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\text{Var}(\widehat{\theta}_2)}{\text{Var}(\widehat{\theta}_1)}.$$

Remark 107. If $\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) > 1$, then $\widehat{\theta}_1$ is a better estimator than $\widehat{\theta}_2$ because it has lower variance.

The next example considers two estimators which are similar to those considered in Example 79 (estimating the sides of a dice).

Example 108 (Example 9.1 in the textbook)**.** Fix $\theta > 0$. Suppose $Y_1, \ldots, Y_n$ is a random sample from the uniform distribution on the interval $(0, \theta)$. Let

$$\widehat{\theta}_1 = 2\overline{Y} \quad \text{and} \quad \widehat{\theta}_2 = \left(\frac{n+1}{n}\right) M,$$

where $M = \max(Y_1, \ldots, Y_n)$.

We'll compute the relative efficiency of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Our computation proceeds through five claims.

Claim 1: $\widehat{\theta}_1$ is unbiased.

Proof of Claim 1. $\mathbb{E}[\widehat{\theta}_1] = \mathbb{E}\left[2\overline{Y}\right] = 2\left(\frac{\theta}{2}\right) = \theta.$ □ Claim

Claim 2: $\text{Var}(\widehat{\theta}_1) = \frac{\theta^2}{3n}$.

Proof of Claim 2. By Proposition 83, the standard deviation of the sample mean $\overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}$ is

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}, \tag{21}$$

where $\sigma^2 = \text{Var}(Y_1)$. Squaring both sides of Eq. (21) implies

$$\text{Var}(\overline{Y}) = \frac{\sigma^2}{n}$$

Therefore

$$\text{Var}(2\overline{Y}) = \frac{4\sigma^2}{n} \tag{22}$$

Moreover, a direct computation will show that since $Y_1 \sim \text{unif}(0, \theta)$, then $\sigma^2 := \text{Var}(Y_i) = \theta^2/12$ (exercise). Plugging this into Eq. (22) implies the statement of the claim. □ Claim

Claim 3: The pdf of M is $f(t) = \frac{n}{\theta^n} t^{n-1} \mathbf{1}_{[0<t<\theta]}$.

Proof of Claim 3. In Example 105, we showed that M has distribution function

$$F(t) = \left(\frac{t}{\theta}\right)^n \mathbf{1}_{[t \in (0,\theta)]}$$

Differentiating F implies the claim. □ Claim

Claim 4: $\widehat{\theta}_2$ is unbiased.

Proof of Claim 4.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}_2\right] &= \left(\frac{n+1}{n}\right)\mathbb{E}\left[M\right] \\
&= \left(\frac{n+1}{n}\right)\int_{-\infty}^{\infty} t f(t)dt \\
&= \frac{n+1}{n}\int_0^{\theta} \frac{n}{\theta^n} t^n dt && \text{by Claim 3} \\
&= \frac{1}{\theta^n}\left[t^{n+1}\right]_0^{\theta} \\
&= \theta.
\end{aligned}$$

□ Claim

Claim 5: $\mathrm{Var}(\widehat{\theta}_2) = \frac{\theta^2}{n(n+2)}$.

Proof of Claim 5. Since $\widehat{\theta}_2 = \left(\frac{n+1}{n}\right) M$, it will be helpful to first compute $\mathbb{E}\left[M^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[M^2\right] &= \int_{-\infty}^{\infty} t^2 f(t)dt && \text{by LOTUS (Theorem 21)} \\
&= ... && \text{plug in formula for } f(t) \text{ and integrate} \\
&= \left(\frac{n}{n+2}\right)\theta^2.
\end{aligned}$$

Therefore

$$\mathbb{E}[\widehat{\theta}_2^2] = \mathbb{E}\left[\left(\frac{n+1}{n}\right)^2 M^2\right] = \left(\frac{n+1}{n}\right)^2\left(\frac{n}{n+2}\right)\theta^2. \tag{23}$$

We now compute $\mathrm{Var}(\widehat{\theta}_2)$:

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}_2) &= \mathbb{E}[\widehat{\theta}_2^2] - \left(\mathbb{E}[\widehat{\theta}_2]\right)^2 \\
&= \mathbb{E}[\widehat{\theta}_2^2] - \theta^2 && \text{since } \mathbb{E}[\widehat{\theta}_2] = \theta \text{ by Claim 4.}
\end{aligned}$$

Plugging Eq. (23) into the above equation and simplifying implies the statement of the claim. □ Claim

By Claim 2 and Claim 5,

$$\mathrm{eff}(\theta_1, \theta_2) = \frac{\left(\frac{\theta^2}{n(n-2)}\right)}{\left(\frac{\theta^2}{3n}\right)} = \frac{3}{n+2}.$$

The efficiency is less than 1 for all $n \geq 2$, and we conclude that $\widehat{\theta}_2$ has a smaller variance than $\widehat{\theta}_1$, making it a better estimator.

End of Example 108. □

Theorem 109. *If $\widehat{\theta}_n$ is an unbiased esetimator of θ and $\lim_{n\to\infty}\mathrm{Var}(\widehat{\theta}_n) = 0$, then $\widehat{\theta}_n$ is consistent.*

Proof idea. Proof by picture. The technical details are similar to the proof of the weak law of large numbers. □