Introduction to Probability and Statistics
MSBA Program

Javier Peña
Professor of Operations Research
Carnegie Mellon University

Fall, 2023

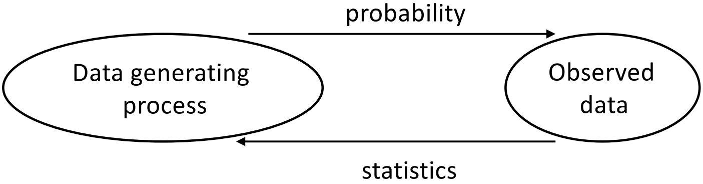
Introduction

Course objectives

- concepts and models to work with uncertainty (probability)
- methods to make intelligent inference from data (statistics)
- foundation for important part of Tepper's MSBA curriculum:

IntroPS → StatsFound → ML1 → ML2

A bird's view of probability and statistics



Probability

Formal mathematical framework to quantify uncertainty.

Statistical inference ("learning")

Use data to infer information about the underlying process that generated the data.

Puzzle: would you take the following gamble?

If at least two students have the same birthday, you pay me one dollar.

Otherwise (all students in this class have different birthdays), I pay you five dollars.

Puzzle: would you take the following gamble?

If at least two students have the same birthday, you pay me one dollar.

Otherwise (all students in this class have different birthdays), I pay you five dollars.

What if I offered you ten dollars (a 10-to-1 gamble)?

Puzzle: would you take the following gamble?

If at least two students have the same birthday, you pay me one dollar.

Otherwise (all students in this class have different birthdays), I pay you five dollars.

What if I offered you ten dollars (a 10-to-1 gamble)?

What if I offered you one hundred dollars (a 100-to-1 gamble)?

Puzzle: testing for COVID-19

There is a *prior* chance that an individual has COVID-19, say 10%.

A COVID-19 test has two main characteristics.
Sensitivity: accuracy of detecting the presence of COVID-19
Specificity: accuracy of detecting the absence of COVID-19

Suppose a test has sensitivity 80% and specificity 90%.

You tested positive. How likely are you to have COVID-19?

About your instructor:

- PhD in Applied Math, Cornell University
- Professor at Tepper for 20+ years
- Research: financial engineering, optimization, quantum computing
- Office: Tepper Quad 4224

Course material

- J. Peña's slides and related files (available in canvas)
- Main reference:
  Business Statistics: Communicating with Numbers, 4th edition, Jaggia and Kelly. Older editions are ok.
- Additional reference for technical depth:
  All of Statistics: A Concise Course in Statistical Inference, Wasserman. Electronic (pdf) version available at the CMU library.

Some administrative issues

Office hours (J. Peña):

Thursday and Friday 9–10am, TEP 4224.

Teaching Assistant:

Arnav Sood
Check canvas for TA office hours info.

Coursework:

Problem sets, quizzes, little tests, and final exam.
Please read syllabus for details.

Class format

- In-class quizzes will provide a way for you to be actively engaged and for me to get valuable feedback.
- Bring your laptop (smartphone with canvas app might do).
- Please be polite and professional. This creates a pleasant and friendly atmosphere. It also builds valuable social capital.
- Please be punctual and refrain from activities that may distract other students.

Course outline

- Core probability concepts
- Random variables and probability distributions
- Convergence of random variables: LLN, CLT
- Point Estimation & Confidence Sets
- Hypothesis Testing
- Regression

Intro to Probability and Statistics, Week 1

This week:

- Sample space, events, probability
- Conditional probability
- Random variables

Sample space, events, probability

Sample space and events

Sample space:

Set all possible outcomes of a random experiment.

It is common to use Ω to denote sample space and $\omega \in \Omega$ to denote a possible outcome.
Less technical textbooks often write S for sample space.

Event:

Subset of the sample space.

It is common to use capital roman letters $A, B, C, ...$ to denote events.

Sample space and events (examples)

Sample spaces

- flip a coin twice: $\Omega = \{HH, HT, TH, TT\}$
- roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- the birthdays of k randomly selected people:

$$\Omega = \{(\omega_1, \cdots, \omega_k) : \omega_i \in \{1, \ldots, 365\}\}.$$

- flip a coin forever:

$$\Omega = \{(\omega_1, \omega_2, \cdots) : \omega_i \in \{H, T\}\}.$$

- spot where the tennis ball lands when I serve
- amount of time it takes to drive to the airport

Sample space and events (examples)

Events

- heads on the first coin flip: $B = \{HH, HT\}$
- roll an even number: $A = \{2, 4, 6\}$
- all birthdays are different
- no H in the first 10 flips
- my serve landed in the correct square of the court
- get to the airport in 40 minutes or less

Set operations

Suppose A and B are events in a sample space Ω.

New events:

- **union:** $A \cup B =$ "either A or B or both"

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$

- **intersection:** $A \cap B =$ "both A and B"

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$

- **complement:** $A^c =$ "not A". Sometimes write $\overline{A}$ for A^c

$$A^c = \{\omega \in \Omega : \omega \notin A\}.$$

Two events A and B are **mutually exclusive** if $A \cap B = \emptyset$.

Set operations (infinite case)

Suppose $A_1, A_2, \ldots$ are events in a sample space Ω.

New events:

- **union:**

$$\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\}.$$

- **intersection:**

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}.$$

The events $A_1, A_2, \ldots$ are **mutually exclusive** if $A_i \cap A_j = \emptyset$ whenever $i \neq j$.

Probability measure

Suppose Ω is a sample space. A **probability measure** assigns a number $\mathbb{P}(A)$ to each event $A \subseteq \Omega$ such that

- $\mathbb{P}(A) \geq 0$ for every event A
- $\mathbb{P}(\Omega) = 1$
- If $A_1, A_2, \ldots$ are mutually exclusive then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Technical nuance: the third rule is related to the fact that Ω could be infinite. In that case it may only be possible to assign probabilities to a collection of subsets of Ω called a σ-field.

Probability on finite sample spaces

Consider the case when Ω is finite.

In this case the third probability rule can be simplified:

- If A, B are mutually exclusive events then

$$P(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

If each outcome in a finite sample space Ω is equally likely then for every event $A \subseteq \Omega$ we have

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

where $|A| =$ size of $A =$ number of elements in A.

Properties of probability measures

Suppose A and B are events and $\mathbb{P}$ is a probability measure.

We will use the following properties many times:

1. Complement rule: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. Total probability rule: $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
4. $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$

Examples

- Flip a coin twice. What is the probability that I get tails at least once?

$$\begin{aligned}\mathbb{P}(\text{tails at least once}) &= 1 - \mathbb{P}(\text{no tails}) \\ &= 1 - \frac{1}{4} \\ &= \frac{3}{4}.\end{aligned}$$

- Roll two dice. What is the probability that the two rolls are different?

$$\begin{aligned}\mathbb{P}(\text{two rolls are different}) &= 1 - \mathbb{P}(\text{two rolls are the same}) \\ &= 1 - \frac{6}{36} \\ &= \frac{5}{6}.\end{aligned}$$

Counting rules

Rules to determine the size of some sample spaces and events.
(Like in previous examples.)

- *Sequences*
  Number of k-sequences from a set of size n:

$$n^k$$

- *Permutations*
  Number of k-sequences without repetition from a set of size n:

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

- *Product of sample spaces*
  Number of k-sequences from k sets of sizes $n_1, \ldots, n_k$:

$$n_1 \cdot n_2 \cdots n_k$$

Conditional probability

Conditional probability

Adjust *prior* likelihood of an event in light of some evidence.

Suppose A and B are events in a sample space Ω. Suppose A occurs, how can we revise $\mathbb{P}(B)$?

The **conditional probability** of B given A is

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

Example

- Draw a random card from a 52-card poker deck. How likely is it to be the ace of diamonds?
- Suppose the drawn card is an ace. How likely it is to be the ace of diamonds?

Probability tables

Suppose A and B are events in a sample space Ω.

Recall: $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$

Summarize probabilities involving A and B in a table:

	B	B^c	
A	$\mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B^c)$	$\mathbb{P}(A)$
A^c	$\mathbb{P}(A^c \cap B)$	$\mathbb{P}(A^c \cap B^c)$	$\mathbb{P}(A^c)$
	$\mathbb{P}(B)$	$\mathbb{P}(B^c)$	1

Probability tables

Suppose A and B are events and consider the probability table

	B	B^c	
A	$\mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B^c)$	$\mathbb{P}(A)$
A^c	$\mathbb{P}(A^c \cap B)$	$\mathbb{P}(A^c \cap B^c)$	$\mathbb{P}(A^c)$
	$\mathbb{P}(B)$	$\mathbb{P}(B^c)$	1

Observe

- Rows and columns add up. For instance in the first row

$$\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A).$$

Likewise for the other rows and columns.

- As a consequence

$$\mathbb{P}(B|A) + \mathbb{P}(B^c|A) = 1.$$

Statistical independence

Suppose A and B are events in a sample space Ω.

We say that A and B are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

If $\mathbb{P}(A) > 0$ then the above condition is equivalent to

$$\mathbb{P}(B) = \mathbb{P}(B|A).$$

Example

Draw a random card from a 52-card poker deck.
$A :=$ drawn card is an ace, $B :=$drawn card is a diamond,
$C :=$ drawn card is the ace of diamonds.
Observe that A and B are independent but A and C are not.

Intuition for independence

Probability of one event is not affected by the other.

Bayes' Theorem

Adjust *prior* likelihood of an event in light of some evidence.

In probability terms:

Suppose $\mathbb{P}(A)$ and $\mathbb{P}(B|A), \mathbb{P}(B|A^c)$ are known. What is $\mathbb{P}(A|B)$?

Bayes' Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B|A) \cdot \mathbb{P}(A) + \mathbb{P}(B|A^c) \cdot \mathbb{P}(A^c)}.$$

The second formula follows from:

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) + \mathbb{P}(B|A^c) \cdot \mathbb{P}(A^c).$$

Example (COVID-19 testing)

A COVID-19 test has two main characteristics.
Sensitivity: accuracy of detecting the presence of COVID-19
Specificity: accuracy of detecting the absence of COVID-19

Without a test, the *baseline* or *prior* probability that an individual has COVID-19 is 10%.

Suppose a test has sensitivity 80% and specificity 90%.

You tested positive. How likely are you to have COVID-19?

COVID-19 testing

Main events:

$A =$ individual has COVID-19.
$B =$ individual is tested positive.

We have

$$\mathbb{P}(A) = 0.1,\ \mathbb{P}(B|A) = 0.8,\ \mathbb{P}(B^c|A^c) = 0.9.$$

and we want $P(A|B)$.

Observe:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 0.9 \text{ and } \mathbb{P}(B|A^c) = 1 - \mathbb{P}(B^c|A^c) = 0.1.$$

Hence

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\cdot\mathbb{P}(A)}{\mathbb{P}(B|A)\cdot\mathbb{P}(A) + \mathbb{P}(B|A^c)\cdot\mathbb{P}(A^c)} \\ &= \frac{0.8\cdot 0.1}{0.8\cdot 0.1 + 0.1\cdot 0.9} = 0.47.\end{aligned}$$

Optional: two neat applications

Monty Hall

There are three doors. A prize is placed at random behind one of the doors. You pick a door. The game host opens one of the remaining doors and reveals that it is empty. You are offered to switch to the thus far unopened door. Should you?

Gambler's Ruin

A gambler enters a casino with capital $a > 0$ and bets repeatedly until his fortune reaches $c > a$ or his funds are exhausted. Both a and c are integer numbers.

In each bet the gambler wins one dollar with probability p and loses one dollar with probability $q = 1 - p$.

What is the probability that the gambler will succeed in achieving his goal c?

Solution to Monty Hall

Let $\omega_1 \in \{1, 2, 3\}$ be the door where the prize is and $\omega_2 \in \{1, 2, 3\}$ be the door that the host opens. The sample space is

$$\{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)\}.$$

If you pick a door (say door 1) and do not switch, you win in two of the six possible outcomes. if you switch, you win in four of the six possible outcomes.

There are several other ways of solving this problem.

Solution to Gambler's Ruin

Let $s_c(a)$ denote the probability of success. Observe that $s_c(c) = 1, s_c(0) = 0$, and

$$s_c(a) = ps_c(a+1) + qs_c(a-1) \text{ for } 1 \leq a \leq c-1.$$

Why?

To solve the above, let $r := q/p$ and $d(a) := s_c(a+1) - s_c(a)$. Then for $a = 0, 1, \ldots, c-1$ we have $s_c(a+1) - s_c(a) = d(a) = rd(a-1)$ and consequently

$$s_c(a) = s_c(a) - s_c(0) = \sum_{i=0}^{a-1} s_c(i+1) - s_c(i) = \sum_{i=0}^{a-1} r^i d(0).$$

After some algebraic work we get

$$s_c(a) = \begin{cases} \frac{a}{c} & \text{if } r = \frac{q}{p} = 1 \\ \frac{r^a - 1}{r^c - 1} & \text{if } r = \frac{q}{p} \neq 1. \end{cases}$$

Examples

Red-and-black: a type of bet in the roulette wheel. There are 38 spaces, 18 are red, 18 are black, and 2 are green. Gambler can bet on red or black. Thus $p = 18/38$ and so $r = 20/18$.

Initial capital $a =$\$900 and goal $c =$ \$1,000

If $p = 1/2$, the chance of success is 0.9.
At red-and-black $r = 20/18$, chance of success ≈ 0.00002656.

Initial capital $a =$\$100 and goal $c =$ \$1,000

If $p = 1/2$, the chance of success is 0.1.
At red-and-black $r = 20/18$, chance of success $\approx 6.6 \times 10^{-42}$.

Random variables

Random variable

Informal definition:

A function that assigns a numerical value to the outcome of a random experiment.

Precise definition:

A random variable X is a real-valued function defined on an underlying sample space Ω, that is,

$$X : \Omega \to \mathbb{R}.$$

The random variable X assigns a value $X(\omega) \in \mathbb{R}$ to each outcome $\omega \in \Omega$.

Note: most of the time we work directly with random variables and do not mentioned the underlying sample space.

Examples

- roll of a die
- flip a coin n times and count the number of Hs
- amount of rainfall in Pittsburgh next year
- amount of time it will take to drive to the airport
- number of guests that will check in at a hotel tonight
- price of Amazon stock at the end of the year

The distribution of a random variable

Suppose $\mathbb{P}$ is a probability on a sample space Ω and $X : \Omega \to \mathbb{R}$ is a random variable.

Notation

For $x \in \mathbb{R}$ and $A \subseteq \mathbb{R}$ let

$$\begin{aligned} \mathbb{P}(X = x) &:= \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right) \\ \mathbb{P}(X \in A) &:= \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right). \end{aligned}$$

The *distribution of* X is the mapping $A \mapsto \mathbb{P}(X \in A)$ for $A \subseteq \mathbb{R}$.

Cumulative distribution function (cdf)

Suppose X is a random variable. The cumulative distribution function (cdf) of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

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

Common convention: Use an uppercase letter X for a random variable and lowercase letter x for some particular value that X may take.

Discrete random variables

A random variable X is discrete if it takes countably many values. That is, if its set of possible values can be listed:

$$x_1, x_2, \ldots$$

Probability mass function (pmf)

Suppose X is a discrete random variable. The probability mass function (pmf) of X is defined by

$$f_X(x) = \mathbb{P}(X = x).$$

When the random variable X is clear from the context, it is common to write f for f_X and F for F_X.

Example

X = roll of a die.

x	$\mathbb{P}(X = x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

Example

Toss a coin twice and let X = number of Hs.

x	$\mathbb{P}(X = x)$
0	$1/4$
1	$1/2$
2	$1/4$

The distribution of a discrete random variable

Observe

- Both pmf and cdf are functions of x:

$$\mathbb{P}(X = x) \quad \text{and} \quad \mathbb{P}(X \leq x)$$

- They are closely related: Suppose X is discrete with values

$$x_1 < x_2 < \cdots$$

If $x_k \leq x < x_{k+1}$ then

$$\mathbb{P}(X \leq x) = \sum_{i=1}^{k} \mathbb{P}(X = x_i).$$

Expectation

Suppose X is a discrete random variable with values $x_1, x_2, \dots$ and

$$p_i := \mathbb{P}(X = x_i),\ i = 1, 2, \dots$$

The **expected value** (or mean) of X is

$$\mathbb{E}(X) = p_1 x_1 + p_2 x_2 + \cdots = \sum_x x \cdot \mathbb{P}(X = x)$$

It is common to write μ or μ_X for $\mathbb{E}(X)$.

Example

Toss a coin twice and let $X =$ number of Hs.

$$\mathbb{E}(X) = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1.$$

Functions of random variables

Suppose X is a discrete random variable and $r : \mathbb{R} \to \mathbb{R}$ is some function.

Then $r(X)$ is also a random variable and

$$\mathbb{E}(r(X)) = \sum_x r(x) \cdot \mathbb{P}(X = x).$$

Example

Toss a coin twice and let $X =$ number of Hs.

$$\mathbb{E}(X^2) = 0^2 \cdot 1/4 + 1^2 \cdot 1/2 + 2^2 \cdot 1/4 = 3/2,$$

$$\mathbb{E}(2^X) = 2^0 \cdot 1/4 + 2^1 \cdot 1/2 + 2^2 \cdot 1/4 = 9/4.$$

Variance

Suppose X is a discrete random variable with expected value μ_X.

The **variance** of X is

$$\mathsf{var}(X) = \mathbb{E}((X-\mu_X)^2) = \sum_x (x-\mu_X)^2 \cdot \mathbb{P}(X=x).$$

It is common to write σ^2 or σ_X^2 for var(X).

The **standard deviation** of X is $\sigma_X = \sqrt{\mathsf{var}(X)}$

Example

Toss a coin twice and let $X =$ number of Hs.

$$\sigma^2 = (0-1)^2 \cdot 1/4 + (1-1)^2 \cdot 1/2 + (2-1)^2 \cdot 1/4 = 1/2.$$

Properties of expectation and variance

Suppose X is a random variable and a, b are numbers.

Then

- $\mathbb{E}(aX) = a \cdot \mathbb{E}(X)$
- $\text{var}(aX) = a^2 \cdot \text{var}(X)$
- $\mathbb{E}(aX + b) = a \cdot \mathbb{E}(X) + b$
- $\text{var}(aX + b) = a^2 \cdot \text{var}(X)$

The variance can also be written as

$$\text{var}(X) = \mathbb{E}(X^2) - \mu_X^2.$$

Why?

Joint probability distribution

Suppose X and Y are discrete random variables.

The **joint probability mass function** of X and Y is

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

Once again, it is common to write $f(x,y)$ for $f_{X,Y}(x,y)$ if X, Y are clear from the context.

If $r : \mathbb{R}^2 \to \mathbb{R}$ then $r(X,Y)$ is a random variable and

$$\mathbb{E}(r(X,Y)) = \sum_x \sum_y r(x,y) \cdot \mathbb{P}(X = x, Y = y).$$

To be precise, suppose $X, Y : \Omega \to \mathbb{R}$ are discrete random variables. Then for $x, y \in \mathbb{R}$

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\})$$

Joint probability tables

The joint probability mass function of two discrete random variables can be displayed in a probability table.

Example

Flip a coin three times.

$X =$ number of Hs in first two flips,
$Y =$ number of Hs in last two flips.

Joint probability table

	$Y = 0$	$Y = 1$	$Y = 2$	X: marginal
$X = 0$	$1/8$	$1/8$	0	$1/4$
$X = 1$	$1/8$	$1/4$	$1/8$	$1/2$
$X = 2$	0	$1/8$	$1/8$	$1/4$
Y: marginal	$1/4$	$1/2$	$1/4$	1

In this example: $\mathbb{E}(X \cdot Y) = \frac{5}{4}$.

Another example

Flip a coin four times.

$X =$ number of Hs in first two flips,
$Z =$ number of Hs in last two flips.

Joint probability table

	$Z = 0$	$Z = 1$	$Z = 2$	X: marginal
$X = 0$	$1/16$	$1/8$	$1/16$	$1/4$
$X = 1$	$1/8$	$1/4$	$1/8$	$1/2$
$X = 2$	$1/16$	$1/8$	$1/16$	$1/4$
Z: marginal	$1/4$	$1/2$	$1/4$	1

In this example: $\mathbb{E}(X \cdot Z) = 1$.

Covariance

Suppose X, Y are random variables with expected values μ_X, μ_Y.

The **covariance** of X and Y is

$$\begin{aligned} \mathsf{cov}(X,Y) &= \mathbb{E}((X-\mu_X)(Y-\mu_Y)) \\ &= \sum_x \sum_y (x-\mu_X)(y-\mu_Y) \cdot \mathbb{P}(X=x, Y=y). \end{aligned}$$

It is common to write $\sigma_{X,Y}$ for $\mathsf{cov}(X,Y)$.

The covariance can also be written as

$$\mathsf{cov}(X,Y) = \mathbb{E}(X \cdot Y) - \mu_X \cdot \mu_Y.$$

Example

For the X, Y, Z variables above (coin flips examples) we have

$$\mu_X = \mu_Y = \mu_Z = 1, \ \ \mathsf{var}(X) = \mathsf{var}(Y) = \mathsf{var}(Z) = \frac{1}{2}.$$

and

$$\mathsf{cov}(X, Y) = \frac{1}{4}, \ \mathsf{cov}(X, Z) = 0.$$

More about expectation and variance

Assume X, Y are random variables and a, b are numbers.

Then

- $\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y)$
- $\mathsf{var}(aX + bY) = a^2 \cdot \mathsf{var}(X) + b^2 \cdot \mathsf{var}(Y) + 2ab \cdot \mathsf{cov}(X, Y)$

Correlation

Assume X, Y are random variables. The **correlation** of X, Y is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}.$$

Correlation is always between -1 and 1.

It is a measure of linear dependence between X and Y.

Independence of random variables

Assume X and Y are discrete random variables.

The variables X, Y are **independent** if for all x, y we have

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y).$$

Covariance and independence

If X, Y are independent then $\text{cov}(X, Y) = 0$.

Therefore, if X, Y are independent then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

Note: $\text{cov}(X, Y) = 0$ does not imply that X, Y are independent.
Can you construct an example?

Bernoulli trial

Random variable X that takes two values: 1 and 0.

Let $p = \mathbb{P}(X = 1)$ and consequently $1 - p = \mathbb{P}(X = 0)$.

Observe

$$\mathbb{E}(X) = p,\ \mathsf{var}(X) = p(1 - p).$$

In this case we say

X has Bernoulli distribution with probability of success p.

Binomial random variable

Number of successes in n independent Bernoulli trials:

$$X = Y_1 + \cdots + Y_n$$

where

- each Y_i has Bernoulli distribution with probability of success p
- $Y_1, \ldots, Y_n$ are independent

Observe: X can take values $0, 1, \ldots, n$ and

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n.$$

In this case we say

X has binomial distribution with parameters n and p.

Shorthand: $X \sim B(n, p)$.

Properties of the binomial distribution

Assume $X \sim B(n, p)$.

Expectation and variance

$$\mathbb{E}(X) = np, \quad \textsf{var}(X) = np(1-p).$$

Excel functions

$$\mathbb{P}(X = x) = \texttt{BINOM.DIST}(\texttt{x}, \texttt{n}, \texttt{p}, \texttt{0})$$

$$\mathbb{P}(X \leq x) = \texttt{BINOM.DIST}(\texttt{x}, \texttt{n}, \texttt{p}, \texttt{1})$$

R functions

$$\mathbb{P}(X = x) = \texttt{dbinom}(\texttt{x}, \texttt{n}, \texttt{p})$$

$$\mathbb{P}(X \leq x) = \texttt{pbinom}(\texttt{x}, \texttt{n}, \texttt{p})$$

Example

Suppose an airline passenger is a "no-show" with probability 0.1. Suppose the airline sold 20 tickets for an 18-seat plane.

Questions

- What is the probability that exactly 18 passengers show up?
Solution. Let $X =$ number of passengers who show up. Then $X \sim B(20, 0.9)$ and we want

$$\mathbb{P}(X = 18) = \texttt{BINOM.DIST}(18, 20, 0.9, 0) = 0.285$$

- What is the probability that the plane is overbooked? That is, what is the probability that more than 18 passengers show up?
Solution. We want

$$\mathbb{P}(X > 18) = 1 - \mathbb{P}(X \leq 18) = 1 - \texttt{BINOM.DIST}(18, 20, 0.9, 1) = 0.391$$

- What is the probability that the number of passengers who show up is between 15 and 18?
  Solution. We want

$$
\begin{aligned}
\mathbb{P}(15 \leq X \leq 18)& \\
&= \mathbb{P}(X \leq 18) - \mathbb{P}(X \leq 14) \\
&= \texttt{BINOM.DIST}(18, 20, 0.9, 1) - \texttt{BINOM.DIST}(14, 20, 0.9, 1) \\
&= 0.59699
\end{aligned}
$$

Example (continued)

Suppose that if a plane is overbooked, the airline has to pay a \$200 fee to each bumped passenger. Suppose the airline sold 20 tickets for an 18-seat plane.

How much (on average) does the airline pay in overbooking fees? That is, what is the expected value of overbooking fees?

Solution. Let $Z =$ overbooking fee. Then Z can take values $0, 200, 400$ with the following probabilities:

$$\begin{aligned}
\mathbb{P}(Z = 0) &= \mathbb{P}(X \leq 18) = \texttt{BINOM.DIST}(18, 20, 0.9, 1) = 0.6083 \\
\mathbb{P}(Z = 200) &= \mathbb{P}(X = 19) = \texttt{BINOM.DIST}(19, 20, 0.9, 0) = 0.2701 \\
\mathbb{P}(Z = 400) &= \mathbb{P}(X = 20) = \texttt{BINOM.DIST}(20, 20, 0.9, 0) = 0.1215.
\end{aligned}$$

We want

$$\mathbb{E}(Z) = 0 \cdot 0.6083 + 200 \cdot 0.2701 + 400 \cdot 0.1215 = 102.66.$$

Optional: an application of the binomial distribution

Tranching

A bank has extended four risky loans (of the same size).

Each borrower defaults with probability 0.1.

The bank builds “tranches” A, B, C, D for investors by pooling together and structuring the loan payments as follows:

- The payments are made in order of priority: first A, then B, then C, then D.
- This means that the most *senior* tranche (tranche A) has top priority in case one or more borrowers default.
- Likewise the most *junior* tranche (tranche D) has the least priority in case one or more borrowers default.

Assume the borrowers’ default events are independent.

What is the probability of default of each tranche?

Let X = number of borrowers that default. Then $X \sim B(4, 0.1)$.

Default event of each tranche can be stated in terms of X:

(a) $\mathbb{P}(\text{tranche A defaults}) = \mathbb{P}(X = 4) =$ $\texttt{binom.dist}(4, 4, 0.1, 0) = 0.0001$.

(b) $\mathbb{P}(\text{tranche B defaults}) = \mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X \leq 2) =$ $1 - \texttt{binom.dist}(2, 4, 0.1, 1) = 0.0037$.

(c) $\mathbb{P}(\text{tranche C defaults}) = \mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 1) =$ $1 - \texttt{binom.dist}(1, 4, 0.1, 1) = 0.0523$.

(d) $\mathbb{P}(\text{tranche D defaults}) = \mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) =$ $1 - \texttt{binom.dist}(0, 4, 0.1, 0) = 0.3439$.

Intro to Probability and Statistics, Week 2

Last week

- Sample space, events, probability
- Conditional probability
- Random variables

This week

- Discrete distributions: binomial, geometric, Poisson
- Continuous distributions: uniform, exponential, normal
- If there is time: negative binomial, hypergeometric, t, Gamma

Discrete distributions

Recall: Bernoulli trial

Random variable Y that takes two values: 1 and 0.

Let $p = \mathbb{P}(Y = 1)$ and consequently $1 - p = \mathbb{P}(Y = 0)$.

Observe

$$\mathbb{E}(Y) = p,\ \mathsf{var}(Y) = p(1 - p).$$

In this case we say

Y has Bernoulli distribution with probability of success p.

Recall: binomial random variable

Number of successes in n independent Bernoulli trials:

$$X = Y_1 + \cdots + Y_n$$

where

- each Y_i has Bernoulli distribution with probability of success p
- $Y_1, \ldots, Y_n$ are independent.

Observe: X can take values $0, 1, \ldots, n$ and

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n.$$

In this case we say

X has binomial distribution with parameters n and p.

Shorthand: $X \sim B(n, p)$.

Properties of the binomial distribution

Assume $X \sim B(n, p)$.

Expectation and variance

$$\mathbb{E}(X) = np, \quad \mathsf{var}(X) = np(1-p).$$

Excel functions

$$\mathbb{P}(X = x) = \texttt{BINOM.DIST}(\texttt{x}, \texttt{n}, \texttt{p}, \texttt{0})$$

$$\mathbb{P}(X \leq x) = \texttt{BINOM.DIST}(\texttt{x}, \texttt{n}, \texttt{p}, \texttt{1})$$

R functions

$$\mathbb{P}(X = x) = \texttt{dbinom}(\texttt{x}, \texttt{n}, \texttt{p})$$

$$\mathbb{P}(X \leq x) = \texttt{pbinom}(\texttt{x}, \texttt{n}, \texttt{p})$$

Example

Suppose an airline passenger is a "no-show" with probability 0.1. Suppose the airline sold 20 tickets for an 18-seat plane.

Questions

- What is the probability that exactly 18 passengers show up?
  Solution. Let $X =$ number of passengers who show up. Then $X \sim B(20, 0.9)$ and we want

$$\mathbb{P}(X = 18) = \texttt{BINOM.DIST}(18, 20, 0.9, 0) = 0.285$$

- What is the probability that the plane is overbooked? That is, what is the probability that more than 18 passengers show up?
  Solution. We want

$$\mathbb{P}(X > 18) = 1 - \mathbb{P}(X \leq 18) = 1 - \texttt{BINOM.DIST}(18, 20, 0.9, 1) = 0.391$$

- What is the probability that the number of passengers who show up is between 15 and 18?
  Solution. We want

$$\begin{aligned}
\mathbb{P}(15 \leq & X \leq 18) \\
&= \mathbb{P}(X \leq 18) - \mathbb{P}(X \leq 14) \\
&= \texttt{BINOM.DIST}(18, 20, 0.9, 1) - \texttt{BINOM.DIST}(14, 20, 0.9, 1) \\
&= 0.59699
\end{aligned}$$

Example (continued)

Suppose that if a plane is overbooked, the airline has to pay a \$200 fee to each bumped passenger. Suppose the airline sold 20 tickets for an 18-seat plane.

How much (on average) does the airline pay in overbooking fees? That is, what is the expected value of overbooking fees?

Solution. Let $Z =$ overbooking fee. Then Z can take values $0, 200, 400$ with the following probabilities:

$$\begin{aligned} \mathbb{P}(Z = 0) &= \mathbb{P}(X \leq 18) = \texttt{BINOM.DIST}(18, 20, 0.9, 1) = 0.602 \\ \mathbb{P}(Z = 200) &= \mathbb{P}(X = 19) = \texttt{BINOM.DIST}(19, 20, 0.9, 0) = 0.2701 \\ \mathbb{P}(Z = 400) &= \mathbb{P}(X = 20) = \texttt{BINOM.DIST}(20, 20, 0.9, 0) = 0.1215. \end{aligned}$$

We want

$$\mathbb{E}(Z) = 0 \cdot 0.602 + 200 \cdot 0.2701 + 400 \cdot 0.1215 = 102.66.$$

Geometric distribution

Geometric random variable

Number X of independent Bernoulli trials until first success.

Observe

X can take values $1, 2, \ldots$ and

$$\mathbb{P}(X = x) = (1 - p)^{x-1} p \text{ for } x = 1, 2, \ldots.$$

In this case we say

X has geometric distribution with parameter p.

Variation

Sometimes (e.g. in R) count the number of failures until first success, that is, $Y := X - 1$. The variable Y takes values $0, 1, \ldots$ and

$$\mathbb{P}(Y = y) = (1 - p)^{y} p \text{ for } y = 0, 1, 2, \ldots.$$

Properties of the geometric distribution

Assume X has geometric distribution with parameter p.

Expectation and variance

$$\mathbb{E}(X) = \frac{1}{p}, \quad \mathsf{var}(X) = \frac{1-p}{p^2}.$$

Cumulative distribution

$$\mathbb{P}(X \le x) = 1 - (1-p)^x \ \text{ for } \ x = 1, 2, \ldots.$$

Memoryless property

$$\mathbb{P}(X > x + y \,|\, X > x) = \mathbb{P}(X > y) = (1-p)^y \ \text{ for } \ x, y = 1, 2, \ldots.$$

Example

The probability that the SP500 moves by 1% or more on a single day is 0.1. Suppose the daily moves are independent of each other.

- How likely is the SP500 not to move by 1% or more in each of the next 10 days?
  Solution. Let $X =$ number of days until next large (1% or more) move. We want
  $$\mathbb{P}(X > 10) = 0.9^{10} = 0.348.$$
- What is the expected number of days until the next day when the SP500 moves by 1% or more?
  Solution. We want
  $$\mathbb{E}(X) = \frac{1}{0.1} = 10.$$

Poisson distribution

Model for a counting process = number of occurrences of some event over time or space.

Poisson process

An experiment observed over time or space that satisfies the following properties:

- Occurrences in non-overlapping intervals are independent.
- The probability distribution of occurrences in any interval is the same for all intervals of equal size.
- The expected value of occurrences in any interval is proportional to the interval size.

Poisson random variable

Number of occurrences achieved by a Poisson process in a specified time or space interval.

Poisson distribution

Examples

- Number of visits at a website during the next hour
- Number of arrivals at an ATM machine in a day
- Number of defects in a 50-yard roll of fabric
- Number of leaks in a 1-mile stretch of a pipeline

A random variable X has Poisson distribution with parameter λ if

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!},\ x = 0, 1, 2, \ldots$$

Shorthand: $X \sim \text{Pois}(\lambda)$.

Properties of the Poisson distribution

Assume $X \sim \text{Pois}(\lambda)$.

Expectation and variance

$$\mathbb{E}(X) = \lambda, \quad \text{var}(X) = \lambda$$

Expected number of occurrences during a time interval of length t:

$$\lambda t.$$

Excel functions

$$\mathbb{P}(X = x) = \texttt{POISSON.DIST}(\texttt{x}, \lambda, \texttt{0})$$

$$\mathbb{P}(X \leq x) = \texttt{POISSON.DIST}(\texttt{x}, \lambda, \texttt{1})$$

R functions

$$\mathbb{P}(X = x) = \texttt{dpois}(\texttt{x}, \lambda)$$

$$\mathbb{P}(X \leq x) = \texttt{ppois}(\texttt{x}, \lambda)$$

Example

Suppose that the arrivals at Amazon's website follow a Poisson process with an average rate of 20 customers per minute.

How likely are the following events:

- At most 50 customers arrive during the next two minutes.
  Solution. Let $X =$ number of arrivals per two-min interval. Then $X \sim \text{Pois}(40)$ and we want
  $$\mathbb{P}(X \leq 50) = \texttt{POISSON.DIST}(50, 40, 1) = 0.94737$$

- More than 10 customers arrive during the next 30 seconds.
  Solution. Let $Y =$ number of arrivals per 30-sec interval. Then $Y \sim \text{Pois}(10)$ and we want
  $$\mathbb{P}(Y > 10) = 1 - \mathbb{P}(Y \leq 10) = 1 - \texttt{POISSON.DIST}(10, 10, 1) = 0.4169$$

- The next customer arrives within 10 seconds.
  Solution. Let $Z =$ number of arrivals per 10-sec interval. Then $Z \sim \text{Pois}(20/6)$ and we want
  $$\mathbb{P}(Z \geq 1) = 1 - \mathbb{P}(Z = 0) = 1 - \texttt{POISSON.DIST}(0, 20/6, 0) = 0.96432$$

Neat properties of the binomial and Poisson distributions

Fact

If $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ are independent then

$$X_1 + X_2 \sim B(n_1 + n_2, p).$$

Fact

If $X_1 \sim \textsf{Pois}(\lambda_1)$ and $X_2 \sim \textsf{Pois}(\lambda_2)$ are independent then

$$X_1 + X_2 \sim \textsf{Pois}(\lambda_1 + \lambda_2).$$

Poisson as a limit of the binomial distribution

Suppose $\lambda > 0$ and $X_n \sim B(n, \lambda/n)$. Then X_n is approximately Poisson for large n: for $x = 0, 1, \ldots$

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

A neat application of binomial: tranching

A bank has extended four risky loans (of the same size).

Each borrower defaults with probability 0.1.

The bank builds “tranches” A, B, C, D for investors by pooling together and structuring the loan payments as follows:

- The payments are made in order of priority: first A, then B, then C, then D.
- This means that the most *senior* tranche (tranche A) has top priority in case one or more borrowers default.
- Likewise the most *junior* tranche (tranche D) has the least priority in case one or more borrowers default.

Assume the borrowers’ default events are independent.

What is the probability of default of each tranche?

Let X = number of borrowers that default. Then $X \sim B(4, 0.1)$.

Default event of each tranche can be stated in terms of X:

(a) $\mathbb{P}(\text{tranche A defaults}) = \mathbb{P}(X = 4) =$ $\texttt{binom.dist}(4, 4, 0.1, 0) = 0.0001.$

(b) $\mathbb{P}(\text{tranche B defaults}) = \mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X \leq 2) =$ $1 - \texttt{binom.dist}(2, 4, 0.1, 1) = 0.0037.$

(c) $\mathbb{P}(\text{tranche C defaults}) = \mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 1) =$ $1 - \texttt{binom.dist}(1, 4, 0.1, 1) = 0.0523.$

(d) $\mathbb{P}(\text{tranche D defaults}) = \mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) =$ $1 - \texttt{binom.dist}(0, 4, 0.1, 0) = 0.3439.$

If there is time: negative binomial and hypergeometric distributions

Negative binomial distribution

Similar to the geometric distribution.

Suppose an infinite sequence of independent Bernoulli trials with probability of success p.

Negative binomial random variable

Number X of failures that occur before a specified number of successes, say r, occur. In other words, X is the number of failures before the r-th success.

Observe:

X can take values $0, 1, 2, \dots$ and

$$\mathbb{P}(X = x) = \binom{r + x - 1}{x} p^r (1 - p)^x \text{ for } x = 0, 1, 2, \dots.$$

Shorthand: $X \sim NB(r, p)$.

Properties of the negative binomial distribution

Assume $X \sim NB(r, p)$.

Expectation and variance

$$\mathbb{E}(X) = \frac{r(1-p)}{p}, \quad \mathsf{var}(X) = \frac{r(1-p)}{p^2}.$$

Excel functions

$$\mathbb{P}(X = x) = \texttt{NEGBINOM.DIST}(\texttt{x}, \texttt{r}, \texttt{p}, \texttt{0})$$

$$\mathbb{P}(X \leq x) = \texttt{NEGBINOM.DIST}(\texttt{x}, \texttt{r}, \texttt{p}, \texttt{1})$$

R functions

$$\mathbb{P}(X = x) = \texttt{dnbinom}(\texttt{x}, \texttt{r}, \texttt{p})$$

$$\mathbb{P}(X \leq x) = \texttt{pnbinom}(\texttt{x}, \texttt{r}, \texttt{p})$$

Hypergeometric distribution

Hypergeometric random variable

Number X of successes if we select n items *without* replacement out of a set of size N items where K of them are successes.

Observe:

X can take values from $\max(0, n + K - N)$ to $\min(K, n)$ and

$$\mathbb{P}(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

for $\max(0, n + K - N) \leq x \leq \min(K, n)$.

Shorthand: $X \sim \text{Hypergeometric}(N, K, n)$.

Properties of the hypergeometric distribution

Assume $X \sim \text{Hypergeometric}(N, K, n)$.

Expectation and variance

$$\mathbb{E}(X) = \frac{nK}{N}, \quad \mathsf{var}(X) = \frac{nK(N-K)}{N^2} \cdot \frac{N-n}{N-1}.$$

Excel functions

$$\mathbb{P}(X = x) = \mathtt{HYPGEOM.DIST}(\mathtt{x}, \mathtt{n}, \mathtt{K}, \mathtt{N}, \mathtt{0})$$

$$\mathbb{P}(X \leq x) = \mathtt{HYPGEOM.DIST}(\mathtt{x}, \mathtt{n}, \mathtt{K}, \mathtt{N}, \mathtt{1})$$

R functions

$$\mathbb{P}(X = x) = \mathtt{dhyper}(\mathtt{x}, \mathtt{K}, \mathtt{N} - \mathtt{K}, \mathtt{n})$$

$$\mathbb{P}(X \leq x) = \mathtt{phyper}(\mathtt{x}, \mathtt{K}, \mathtt{N} - \mathtt{K}, \mathtt{n})$$

Continuous distributions

Continuous random variable

Informal definition

Random variable that can take *any* numerical value in an interval.

All possible values of a continuous random variable cannot be listed.

Examples

- Exact arrival time of the school bus that is supposed to pick up my son between 7:40am and 8:00am.
- Exact volume of gasoline that a gas station sells in one day.
- Time until the next call arrives at a 1-800 customer service line.

Probability distribution

Recall that the probability mass function (pmf) of a discrete random variable is

$$f_X(x) = \mathbb{P}(X = x).$$

This function is non-zero only for the values that X takes.

By contrast, for a continuous random variable X and a single possible value x we have

$$\mathbb{P}(X = x) = 0.$$

Instead it makes sense to consider

$$\mathbb{P}(a \leq X \leq b).$$

Probability density function

Definition

A random variable X is continuous if there exists a function $f_X : \mathbb{R} \to \mathbb{R}_+$ such that $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for all $a \leq b$

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \int_a^b f_X(x)dx \\ &= \text{ area under graph of } f_X \text{ between } a \text{ and } b.\end{aligned}$$

The function f_X is the probability density function (pdf) of X.

Observe

- Connection between pdf and cdf:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- For small Δx we have $\mathbb{P}(x \leq X \leq x + \Delta x) \approx f_X(x) \cdot \Delta x$.

Uniform distribution

Suppose a random variable X takes values in $[a, b]$ and all values are equally likely.

Density

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Cumulative distribution

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b. \end{cases}$$

In this case say

X is uniformly distributed in $[a, b]$. Shorthand: $X \sim U(a, b)$.

Example

A gas station has a 1500-gallon tank that is filled in the morning.

Previous experience indicates that any demand between 0 and 1500 gallons per day is equally likely.

Questions

- Probability of selling at most 400 gallons in one day?
- Probability of selling between 500 and 800 gallons in one day?
- Probability of selling between 500 and 530 gallons in one day?
- Probability of selling between 500 and 503 gallons in one day?
- Probability of selling exactly 500 gallons in one day?

Exponential distribution

Related to both geometric and Poisson distributions.

Exponential random variable

Length of time until the next occurrence of a Poisson process.

Examples

- Time until the next customer arrives
- Time until the next call to an 1-800 number
- Time until next operation failure occurs

We say that a random variable X has exponential distribution with parameter $\lambda > 0$ if its range of values is $[0, \infty)$ and it has density

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \lambda e^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

Shorthand: $X \sim \text{Exp}(\lambda)$.

Properties of the exponential distribution

Assume $X \sim \text{Exp}(\lambda)$.

Cumulative distribution

$$F_X(x) = 1 - e^{-\lambda x} \;\; \text{for} \;\; x \geq 0.$$

Excel functions

$$f_X(x) = \texttt{EXPON.DIST}(\texttt{x}, \lambda, \texttt{0}), \;\; F_X(x) = \texttt{EXPON.DIST}(\texttt{x}, \lambda, \texttt{1})$$

R functions

$$f_X(x) = \texttt{dexp}(\texttt{x}, \lambda), \;\; F_X(x) = \texttt{pexp}(\texttt{x}, \lambda)$$

Connection between exponential and Poisson

Exponential	*Poisson*
Time to arrival	Number of arrivals
Avg time between arrivals $= 1/\lambda$	Arrival rate $= \lambda$
Units: time	Units: $1/$time

Observe

Suppose arrivals follow a Poisson process with rate λ and

- $X :=$ time until next arrival
- $Y :=$ number of arrivals in an interval of length t for $t > 0$.

Then $X \sim \exp(\lambda)$, $Y \sim \text{Pois}(\lambda t)$, and

$$\mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = \mathbb{P}(Y = 0) = e^{-\lambda t}.$$

Example

Suppose that on Christmas Eve it is estimated that customers visit Amazon's website at an average rate of 20 customers per minute. Let $X =$ time (in minutes) until the next arrival. Then $X \sim \text{Exp}(20)$. How likely are the following events:

- The next customer arrives within the next 5 seconds.
  Solution. We want
  $$\mathbb{P}(X \leq 1/12) = \texttt{EXPON.DIST}(1/12, 20, 1) = 0.811$$
- No customer arrives during the next 10 seconds.
  Solution. We want
  $$\mathbb{P}(X > 1/6) = 1 - \mathbb{P}(X \leq 1/6) = 1 - \texttt{EXPON.DIST}(1/6, 20, 1) = 0.03567$$
- The next customer arrives between 10 and 15 secs from now.
  Solution. We want
  $$\begin{aligned}\mathbb{P}(1/6 \leq X \leq 1/4) &= \mathbb{P}(X \leq 1/4) - \mathbb{P}(X \leq 1/6) \\ &= \texttt{EXPON.DIST}(1/4, 20, 1) - \texttt{EXPON.DIST}(1/6, 20, 1) \\ &= 0.028936\end{aligned}$$

Another possible approach

Let Y = time (in seconds) until the next arrival. Then $Y \sim \text{Exp}(20/60) = \text{Exp}(1/3)$.

How likely are the following events:

- The next customer arrives within the next 5 seconds.
  Solution. We want
  $$\mathbb{P}(Y \leq 5) = \texttt{EXPON.DIST}(5, 1/3, 1) = 0.811$$
- No customer arrives during the next 10 seconds.
  Solution. We want
  $$\mathbb{P}(Y > 10) = 1 - \mathbb{P}(Y \leq 10) = 1 - \texttt{EXPON.DIST}(10, 1/3, 1) = 0.03567$$
- The next customer arrives between 10 and 15 secs from now.
  Solution. We want
  $$\begin{aligned}\mathbb{P}(10 \leq Y \leq 15) &= \mathbb{P}(Y \leq 15) - \mathbb{P}(Y \leq 10)\\ &= \texttt{EXPON.DIST}(15, 1/3, 1) - \texttt{EXPON.DIST}(10, 1/3, 1)\\ &= 0.028936\end{aligned}$$

Yet another approach (to some of the previous questions)

- Let $Z =$ number of arrivals within the next 5 seconds. Then $Z \sim \text{Pois}(20/12)$.

  Probability that next customer arrives within the next 5 seconds:

  $$\mathbb{P}(Z \geq 1) = 1 - \mathbb{P}(Z = 0) = 1 - \texttt{POISSON.DIST}(0, 20/12, 0) = 0.811$$

- Let $W =$ number of arrivals within the next 10 seconds. Then $W \sim \text{Pois}(20/6)$.

  Probability that no customer arrives within the next 10 seconds:

  $$\mathbb{P}(W = 0) = \texttt{POISSON.DIST}(0, 20/6, 0) = 0.03567$$

Memoryless property of the exponential distribution

If $X \sim \mathrm{Exp}(\lambda)$ for $\lambda > 0$, then

$$\mathbb{P}(X > s + t \,|\, X > s) = \mathbb{P}(X > t) \;\text{ for }\; s, t > 0.$$

In above example: the probability that a customer arrives within the next 5 seconds, given that no customer has arrived in the last 10 seconds is:

$$\mathbb{P}(X > 15 \,|\, X > 10) = \mathbb{P}(X > 5).$$

Expectation and variance

Suppose X is a continuous random variable with density f_X.

The **expected value** (or mean) of X, often denoted μ or μ_X, is

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \text{“average” value of } X\end{aligned}$$

Suppose $r : \mathbb{R} \to \mathbb{R}$ is a function. Then $r(X)$ is also a random variable and

$$\mathbb{E}(r(X)) = \int_{-\infty}^{\infty} r(x) f_X(x) dx.$$

The **variance** of X, often denoted σ^2 or σ_X^2, is

$$\text{var}(X) = \mathbb{E}((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

We will not need to compute any complicated integrals in this course.

Mean and variance for uniform distribution

If $X \sim U(a, b)$ then

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

Mean and variance for exponential distribution

If $X \sim \mathsf{Exp}(\lambda)$ then

$$\mathbb{E}(X) = \frac{1}{\lambda},\ \mathsf{var}(X) = \frac{1}{\lambda^2}.$$

Median and quantiles

Suppose X is a continuous random variable.

The *median* of X is the value m such that

$$\mathbb{P}(X \geq m) = \mathbb{P}(X \leq m) = 1/2.$$

Given $p \in (0,1)$, the p*-th quantile* is the value q such that

$$\mathbb{P}(X \leq q) = p.$$

Equivalently, the p-th quantile is

$$F_X^{-1}(p),$$

where F_X is the cdf (cumulative distribution function) of X.

Special quantiles

Suppose X is a random variable.

- The median is the 0.5-quantile
- The first-quartile, second-quartile, and third-quartile are respectively the 0.25-quantile, 0.5-quantile, and 0.75-quantile
- The first-decile, second-decile,..., ninth-decile are respectively the 0.1-quantile, 0.2-quantile,..., 0.9-quantile
- Similarly for percentiles

Quantiles for general random variables

Suppose X is a random variable with cdf F_X.

Given $p \in (0,1)$, the p-th quantile of X is

$$\min\{x : F_X(x) \geq p\}.$$

If X is a continuous random variable then for all $p \in (0,1)$ we have

$$q = p\text{-th quantile of } X \Leftrightarrow F_X(q) = p \Leftrightarrow q = F_X^{-1}(p).$$

When X is discrete the latter equivalence does not hold.

Example

Suppose $X \sim B(2, 0.5)$. Then the 0.5-quantile of X is 1 but $F_X(1) = 0.75$.

Standard normal distribution

A random variable is **normal** if its density has a peculiar *bell shape.*

A continuous random variable Z is **standard normal** if its range of values is $(-\infty, \infty)$ and it has probability density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

What is so special about normal random variables?

- Central concept in probability and statistics.
- Popular model for many random quantities.
- Fundamental tool for analyzing and drawing conclusions from data.

Properties of the standard normal distribution

Assume Z is a standard normal random variable.

Then

- Z has a bell-shaped distribution
- $\mathbb{E}(Z) = 0, \;\; \text{var}(Z) = 1$
- Symmetric distribution: $\mathbb{P}(Z \geq a) = \mathbb{P}(Z \leq -a)$
- Distribution concentrates around the mean:

$$\begin{aligned} \mathbb{P}(-1 \leq Z \leq 1) &= 0.683 \\ \mathbb{P}(-1.96 \leq Z \leq 1.96) &= 0.95 \\ \mathbb{P}(-2.576 \leq Z \leq 2.576) &= 0.99 \end{aligned}$$

Properties of the standard normal distribution

Assume Z is a standard normal random variable.

Cumulative distribution

$$F(z) = \mathbb{P}(Z \leq z).$$

Excel functions

$$f(z) = \texttt{NORM.S.DIST}(\texttt{z}, \texttt{0}), \quad F(z) = \texttt{NORM.S.DIST}(\texttt{z}, \texttt{1})$$

R functions

$$f(z) = \texttt{dnorm}(\texttt{z}), \quad F(z) = \texttt{pnorm}(\texttt{z})$$

Normal distribution

A random variable X is **normal** if its range of values is $(-\infty, \infty)$ and it has probability density

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

for some parameters μ, σ.

Shorthand: $X \sim N(\mu, \sigma^2)$.

Properties of the normal distribution

Assume $X \sim N(\mu, \sigma^2)$.

- X has a bell-shaped distribution
- $\mathbb{E}(X) = \mu, \;$ var$(X) = \sigma^2$
- Symmetric around μ: $\mathbb{P}(X \geq \mu + a) = \mathbb{P}(X \leq \mu - a)$
- Distribution concentrates around the mean:

$$\begin{aligned} \mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) &= 0.683 \\ \mathbb{P}(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) &= 0.95 \\ \mathbb{P}(\mu - 2.576\sigma \leq X \leq \mu + 2.576\sigma) &= 0.99 \end{aligned}$$

Properties of the normal distribution

Assume $X \sim N(\mu, \sigma^2)$.

Excel functions

$$f(x) = \texttt{NORM.DIST}(\texttt{x}, \mu, \sigma, \texttt{0}), \quad F(x) = \texttt{NORM.DIST}(\texttt{x}, \mu, \sigma, \texttt{1})$$

R functions

$$f(x) = \texttt{dnorm}(\texttt{x}, \mu, \sigma), \quad F(x) = \texttt{pnorm}(\texttt{x}, \mu, \sigma)$$

Example

Suppose the annual return on the SP500 is a normal random variable with mean 10% and standard deviation 15%.

Let $X =$ annual return of SP500. Then $X \sim N(0.10, 0.15^2)$

Find the probability of each of the following events.

- Return is less than 0% (loss): we want

$$\mathbb{P}(X \leq 0) = \texttt{NORM.DIST}(0, 0.10, 0.15, 1) = 0.2524$$

- Return is less than -20% (bad year): we want

$$\mathbb{P}(X \leq -0.20) = \texttt{NORM.DIST}(-0.20, 0.10, 0.15, 1) = 0.0227$$

- Return is more than 40% (a "very good" year): we want

$$\begin{aligned}\mathbb{P}(X \geq 0.40) &= 1 - \mathbb{P}(X \leq 0.40) \\ &= 1 - \texttt{NORM.DIST}(0.40, 0.10, 0.15, 1) \\ &= 0.0227\end{aligned}$$

Example (continued)

- Return is between -5% and 25% (a “typical year”): we want

$$
\begin{aligned}
&\mathbb{P}(-0.05 \leq X \leq 0.25) = \mathbb{P}(X \leq 0.25) - \mathbb{P}(X \leq -0.05) \\
&\quad = \texttt{NORM.DIST}(\texttt{0.25}, \texttt{0.10}, \texttt{0.15}, \texttt{1}) - \texttt{NORM.DIST}(-\texttt{0.05}, \texttt{0.10}, \texttt{0.15}, \texttt{1}) \\
&\quad = 0.682689.
\end{aligned}
$$

Summary of most popular probability distributions

Discrete

Distribution	p.m.f. $\mathbb{P}(X = x)$	$\mathbb{E}(X)$	var(X)
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, \ldots, n$	np	$np(1-p)$
Geometric	$(1-p)^{x-1} p$ for $x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$	λ	λ

Continuous

Distribution	p.d.f. $f(x)$	$\mathbb{E}(X)$	var(X)
Uniform	$\frac{1}{b-a}$ for $x \in [a, b]$ and 0 otherwise	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 otherwise	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

If there is time: t and Gamma distributions

The t-distribution

Symmetric and bell shaped like the normal distribution but with thicker tails.

Let $\nu > 0$. A continuous random variable has t-distribution with ν degrees of freedom if it has density

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is the function

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt.$$

The above function satisfies $\Gamma(x+1) = x\Gamma(x)$ for all $x > 0$. Also, $\Gamma(n+1) = n!$ for $n = 1, 2, \ldots$.

Properties of the t-distribution

Suppose X has t-distribution with ν degrees of freedom.
Shorthand: $X \sim t_\nu$.

Expectation

$$\mathbb{E}(X) = \begin{cases} 0 & \text{for } \nu > 1 \\ \text{undefined} & \text{for } 0 < \nu \leq 1 \end{cases}$$

Variance

$$\text{var}(X) = \begin{cases} \frac{\nu}{\nu-2} & \text{for } \nu > 2 \\ \infty & \text{for } 1 < \nu \leq 2 \\ \text{undefined} & \text{for } 0 < \nu \leq 1 \end{cases}$$

R functions

$$f(x) = \texttt{dt}(\texttt{x}, \nu),\ F(x) = \texttt{pt}(\texttt{x}, \nu)$$

Excel functions

$$f(x) = \texttt{T.DIST}(\texttt{x}, \nu, \texttt{0}),\ F(x) = \texttt{T.DIST}(\texttt{x}, \nu, \texttt{1})$$

Special cases of the t-distribution

As $\nu \to \infty$ the t-distribution approaches the standard normal distribution:

$$\lim_{\nu\to\infty} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

When $\nu = 1$ we get the *Cauchy* distribution

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

The expectation of the Cauchy distribution is undefined due to the size of its tails.

Gamma distribution

Suppose $\alpha, \beta > 0$. A continuous random variable has Gamma distribution with parameters α and β if it has density

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

Expectation and variance

Suppose X has Gamma distribution with parameters α and β. Then

$$\mathbb{E}(X) = \frac{\alpha}{\beta},\ \text{var}(X) = \frac{\alpha}{\beta^2}.$$

This distribution has "shape" $= \alpha$, "rate" $= \beta$, and "scale" $= 1/\beta$.

R functions

$$f(x) = \texttt{dgamma}(\texttt{x}, \alpha, \beta),\ F(x) = \texttt{pgamma}(\texttt{x}, \alpha, \beta)$$

Excel functions

$$f(x) = \texttt{GAMMA.DIST}(\texttt{x}, \alpha, \beta, \texttt{0}),\ F(x) = \texttt{GAMMA.DIST}(\texttt{x}, \alpha, \beta, \texttt{1})$$

Special cases of the Gamma distribution

When $\alpha = 1, \beta = \lambda$ we get Exp(λ).

When $\alpha = m/2, \beta = 1/2$ we get the χ^2 *distribution with* m *degrees of freedom.*

Interesting connection with other distributions

- Suppose $X_1, \ldots, X_m$ are iid standard normal. Then $X_1^2 + \cdots + X_m^2$ has χ^2 distribution with m degrees of freedom.
- Suppose Y, Z are independent, Y has χ^2 distribution with m degrees of freedom, and Z is standard normal. Then

$$\frac{Z}{\sqrt{Y/m}}$$

has t-distribution with m degrees of freedom.

Intro to Prob and Stats, Week 3

Last time

- Continuous random variables
- Continuous distributions: uniform, exponential
- Median and quantiles

Today

- The normal distribution
- Linear transformations of a random variable
- Joint distributions
- If there is time: Law of Large Numbers and Central Limit Theorem

Recap: density, expectation, variance

A continuous random variable X has *probability density function* $f(x)$ if for all $a < b$

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \int_a^b f(x)dx \\ &= \text{ area under graph of } f \text{ between } a \text{ and } b.\end{aligned}$$

The **expected value** (or mean) of X, often denoted μ or μ_X, is

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \text{ ``average'' value of } X\end{aligned}$$

The **variance** of X, often denoted σ^2 or σ_X^2, is

$$\mathsf{var}(X) = \mathbb{E}((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx.$$

Recap: uniform and exponential distributions

Uniform

Suppose $a < b$. We say that $X \sim U(a, b)$ if its density is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Exponential

Suppose $\lambda > 0$. We say that $X \sim \mathrm{Exp}(\lambda)$ if its density is

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ \lambda e^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

Mean and variance for uniform distribution

If $X \sim U(a,b)$ then

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

Mean and variance for exponential distribution

If $X \sim \text{Exp}(\lambda)$ then

$$\mathbb{E}(X) = \frac{1}{\lambda}, \text{ var}(X) = \frac{1}{\lambda^2}.$$

Recap: memoryless property

If $X \sim \text{Exp}(\lambda)$ for $\lambda > 0$, then

$$\mathbb{P}(X > s + t \,|\, X > s) = \mathbb{P}(X > t) \text{ for } s, t > 0.$$

Important

If X has a different distribution then it is not necessarily memoryless.

Example

Suppose $X \sim U(0, 10)$. Then

$$\mathbb{P}(X > 4) = \frac{10 - 4}{10} = 0.6$$

and

$$\mathbb{P}(X > 10 \,|\, X > 6) = 0 \neq \mathbb{P}(X > 4).$$

Recap: median and quantiles

Suppose X is a continuous random variable.

The *median* of X is the value m such that

$$\mathbb{P}(X \geq m) = \mathbb{P}(X \leq m) = 1/2.$$

Given $p \in (0,1)$, the p-*quantile* is the value q such that

$$\mathbb{P}(X \leq q) = p.$$

Equivalently, the p-quantile is

$$F_X^{-1}(p),$$

where F_X is the cdf (cumulative distribution function) of X.

Recap: quantiles for general random variables

Suppose X is a random variable with cdf F_X.

Given $p \in (0,1)$, the p-quantile of X is

$$\min\{x : F_X(x) \geq p\}.$$

If X is a continuous random variable then for all $p \in (0,1)$ we have

$$q = p\text{-quantile of } X \Leftrightarrow F_X(q) = p \Leftrightarrow q = F_X^{-1}(p).$$

When X is discrete the latter equivalence does not hold.

Example

Suppose $X \sim B(2, 0.5)$. Then the 0.5-quantile of X is 1 but $F_X(1) = 0.75$.

The normal distribution

Standard normal distribution

A random variable Z is **standard normal** if its range of values is $(-\infty, \infty)$ and it has probability density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

What is it so special about normal random variables?

- Central concept in probability and statistics.
- Popular model in many contexts.
- Fundamental tool for analyzing and drawing conclusions from data.

Properties of the standard normal distribution

Assume Z is a standard normal random variable.

Then

- Z has a bell-shaped distribution
- $\mathbb{E}(Z) = 0, \;\; \text{var}(Z) = 1$
- Symmetric distribution: $\mathbb{P}(Z \geq a) = \mathbb{P}(Z \leq -a)$
- Distribution concentrates around the mean:

$$\begin{aligned} \mathbb{P}(-1 \leq Z \leq 1) &= 0.683 \\ \mathbb{P}(-1.96 \leq Z \leq 1.96) &= 0.95 \\ \mathbb{P}(-2.576 \leq Z \leq 2.576) &= 0.99 \end{aligned}$$

Properties of the standard normal distribution

Assume Z is a standard normal random variable.

Cumulative distribution

$$F(z) = \mathbb{P}(Z \leq z)$$

is not computable “by hand”.

Excel functions

$$f(z) = \texttt{NORM.S.DIST}(\texttt{z}, 0), \quad F(z) = \texttt{NORM.S.DIST}(\texttt{z}, 1)$$

Excel function for quantile

For $p \in [0, 1]$ the p-quantile is

$$F^{-1}(p) = \texttt{NORM.S.INV}(\texttt{p})$$

Normal distribution

A random variable X is **normal** if its range of values is $(-\infty, \infty)$ and it has probability density

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

for some parameters μ, σ.

Shorthand: $X \sim N(\mu, \sigma^2)$.

Properties of the normal distribution

Assume $X \sim N(\mu, \sigma^2)$.

- X has a bell-shaped distribution
- $\mathbb{E}(X) = \mu, \;$ var$(X) = \sigma^2$
- Symmetric around μ: $\mathbb{P}(X \geq \mu + a) = \mathbb{P}(X \leq \mu - a)$
- Distribution concentrates around the mean:

$$\begin{aligned} \mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) &= 0.683 \\ \mathbb{P}(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) &= 0.95 \\ \mathbb{P}(\mu - 2.576\sigma \leq X \leq \mu + 2.576\sigma) &= 0.99 \end{aligned}$$

Properties of the normal distribution

Assume $X \sim N(\mu, \sigma^2)$.

Excel functions

$$f(x) = \texttt{NORM.DIST}(\texttt{x}, \mu, \sigma, \texttt{0}), \quad F(x) = \texttt{NORM.DIST}(\texttt{x}, \mu, \sigma, \texttt{1})$$

For $p \in (0, 1)$ the p-quantile is

$$F^{-1}(p) = \texttt{NORM.INV}(\texttt{p}, \mu, \sigma)$$

R functions

$$f(x) = \texttt{dnorm}(\texttt{x}, \mu, \sigma), \quad F(x) = \texttt{pnorm}(\texttt{x}, \mu, \sigma)$$

and

$$F^{-1}(p) = \texttt{qnorm}(\texttt{p}, \mu, \sigma)$$

Example

Suppose the annual return on the SP500 is a normal random variable with mean 10% and standard deviation 15%.

Let X = annual return of SP500. Then $X \sim N(0.10, 0.15^2)$

Find the probability of each of the following events.

- Return is less than 0% (loss): we want

$$\mathbb{P}(X \leq 0) = \texttt{NORM.DIST}(0, 0.10, 0.15, 1) = 0.2524$$

- Return is less than -20% (bad year): we want

$$\mathbb{P}(X \leq -0.20) = \texttt{NORM.DIST}(-0.20, 0.10, 0.15, 1) = 0.0227$$

- Return is more than 40% (a "very good" year): we want

$$\begin{aligned}\mathbb{P}(X \geq 0.40) &= 1 - \mathbb{P}(X \leq 0.40) \\ &= 1 - \texttt{NORM.DIST}(0.40, 0.10, 0.15, 1) \\ &= 0.0227\end{aligned}$$

Example (continued)

- Return is between -5% and 25% (a "typical year"): we want

$$\begin{aligned}
&\mathbb{P}(-0.05 \leq X \leq 0.25) = \mathbb{P}(X \leq 0.25) - \mathbb{P}(X \leq -0.05) \\
&\quad = \texttt{NORM.DIST}(0.25, 0.10, 0.15, 1) - \texttt{NORM.DIST}(-0.05, 0.10, 0.15, 1) \\
&\quad = 0.682689.
\end{aligned}$$

Determine the value r such that the return is more than r with probability 0.95.

Solution. We want r such that $\mathbb{P}(X > r) = 1 - \mathbb{P}(X \leq r) = 0.95$. So

$$r = \text{0.05-quantile of } X = \texttt{NORM.INV}(0.05, 0.10, 0.15) = -0.1467.$$

Summary of most popular probability distributions

Discrete

Distribution	p.m.f. $\mathbb{P}(X = x)$	$\mathbb{E}(X)$	var(X)
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, \ldots, n$	np	$np(1-p)$
Geometric	$(1-p)^{x-1} p$ for $x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$	λ	λ

Continuous

Distribution	p.d.f. $f(x)$	$\mathbb{E}(X)$	var(X)
Uniform	$\frac{1}{b-a}$ for $x \in [a, b]$ and 0 otherwise	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 otherwise	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Quantile calculations for popular distributions

Excel and R formulas to compute p-th quantile

$$\min\{x : F_X(x) \geq p\}.$$

Distribution	Excel	R
Standard normal	`NORM.S.INV(p)`	`qnorm(p)`
$N(\mu, \sigma^2)$	`NORM.INV`$(\texttt{p}, \mu, \sigma)$	`qnorm`$(\texttt{p}, \mu, \sigma)$
$B(n, \text{prob})$	`BINOM.INV(n, prob, p)`	`qbinom(p,n,prob)`
$U(a, b)$	–	`qunif(p,a,b)`
$\text{Exp}(\lambda)$	–	`qexp`$(\texttt{p}, \lambda)$
$\text{Pois}(\lambda)$	–	`qpois`$(\texttt{p}, \lambda)$
$\ldots$	–	`q...`

R formulas for distributions

Prob mass/density function (pmf/pdf)

`dbinom(...), dnorm(...), etc`

Cumulative distribution function (cdf)

`pbinom(...), pnorm(...), etc`

Quantile

`qbinom(...), qnorm(...), etc`

Random draw

`rbinom(...), rnorm(...), etc`

Excel formulas

for pmf/pdf and cdf: `BINOM.DIST(...), NORM.DIST(...)`
for quantiles: `BINOM.INV(...), NORM.INV(...)`

Equivalence in distribution

Generic notation/terminology

It is common to say "X has distribution F" and write $X \sim F$ to indicate that F is the cumulative distribution function of X. When X has density f sometimes you also see $X \sim f$.

We say that two random variables X and Y are **equal in distribution** if $F_X = F_Y$. In that case we write

$$X \overset{d}{=} Y.$$

Observe

It is possible for two random variables X, Y on a sample space Ω to satisfy both $X \overset{d}{=} Y$ and $X(\omega) \neq Y(\omega)$ for every $\omega \in \Omega$. Can you give an example?

Linear transformations of a random variable

Linear transformation of a random variable

Suppose X is a random variable and a, b are numbers.

The random variable $Y = aX + b$ is a *linear transformation* of X.

Examples

- $X =$ time (in hours), $Y =$ time (in minutes)
- $X =$ revenue (in dollars), $Y =$ revenue (in euros)
- $X =$ temperature (Celsius), $Y =$ temperature (Fahrenheit)

Properties of expectation and variance

Assume X is a random variable and a, b are numbers. Then

$$\mathbb{E}(aX + b) = a \cdot \mathbb{E}(X) + b$$

and

$$\text{var}(Y) = \text{var}(aX + b) = a^2 \cdot \text{var}(X).$$

Standardization

Suppose X is a random variable and $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{var}(X)$.

Then the variable

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

has $\mathbb{E}(Z) = 0$ and $\text{var}(Z) = 1$.

Neat properties of the normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Then for all $a, b \in \mathbb{R}$ we have

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

In particular, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$.

Linear transformations and density functions

Suppose X is a continuous random variable with density f_X and $a, b \in \mathbb{R}$ with $a \neq 0$.

Then the random variable $Y := aX + b$ is continuous with density f_Y defined by

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|}$$

Joint distributions

Joint probability distribution

Suppose X and Y are random variables.

Discrete case (revisited)

When X, Y are discrete, the *joint probability mass function* is

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

Continuous case

The variables X, Y have *continuous joint distribution* if there exists $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ such that for any $A \subseteq \mathbb{R}^2$

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y) dx dy.$$

To be precise

Suppose $X, Y : \Omega \to \mathbb{R}$. Then for $A \subseteq \mathbb{R}^2$

$$\mathbb{P}((X,Y) \in A) = \mathbb{P}(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}).$$

Joint cumulative distribution

Suppose X and Y are random variables. The *joint cumulative distribution* of (X, Y) is

$$F_{X,Y}(x, y) := \mathbb{P}(X \leq x, Y \leq y).$$

Connection between joint cdf and joint pdf

Suppose X and Y have continuous joint distribution with joint pdf f and joint cdf F. Then

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x, y)}{\partial y \partial x}$$

at every (x, y) where these derivatives exist.

Transformations, covariance, correlation

Suppose X, Y have continuous joint distribution with density $f_{X,Y}$. If $r : \mathbb{R}^2 \to \mathbb{R}$ then the expectation of $r(X, Y)$ is

$$\mathbb{E}(r(X,Y)) = \iint_{\mathbb{R}^2} r(x,y) \cdot f_{X,Y}(x,y)dxdy$$

Suppose

$$\mu_X = \mathbb{E}(X),\ \mu_Y = \mathbb{E}(Y).$$

The **covariance** of X, Y (often denoted $\sigma_{X,Y}$) is

$$\mathsf{cov}(X,Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

The **correlation** of X, Y is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}.$$

The correlation is always between -1 and 1.

Special cases

If $Y = X$ then $\mu_Y = \mu_X, \ \sigma_Y = \sigma_X$, and

$$\sigma_{X,Y} = \sigma_X^2 \text{ and } \rho_{X,Y} = 1$$

If $Y = -X$ then $\mu_Y = -\mu_X, \ \sigma_Y = \sigma_X$, and

$$\sigma_{X,Y} = -\sigma_X^2 \text{ and } \rho_{X,Y} = -1$$

Unlike the covariance, correlation is scale invariant

If $a > 0$ and $b > 0$ then

$$\mathsf{cov}(aX, bY) = ab \cdot \mathsf{cov}(X, Y) \text{ and } \rho_{aX,bY} = \rho_{X,Y}$$

Properties of expectation, variance and covariance

Assume X, Y are random variables and a, b are numbers. Then

- $\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y)$
- $\mathsf{var}(aX + bY) = a^2 \cdot \mathsf{var}(X) + b^2 \cdot \mathsf{var}(Y) + 2ab \cdot \mathsf{cov}(X, Y)$

Special case

Suppose X, Y are uncorrelated, that is, $\rho_{X,Y} = 0$. Then

- $\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y)$
- $\mathsf{var}(aX + bY) = a^2 \cdot \mathsf{var}(X) + b^2 \cdot \mathsf{var}(Y)$

Example

Suppose you can invest in the following two mutual funds:

- A *small cap equity* fund with annual expected return 12% and standard deviation 16%
- A *large cap equity* fund with annual expected return 8% and standard deviation 11%.

Suppose the two returns of the funds are uncorrelated ($\rho = 0$). Find the expected return (percentage) and standard deviation (percentage) of a portfolio that invests 50% in small cap and 50% in large cap. The latter is called a "$(50\%, 50\%)$ allocation"

How different are the above expected returns and standard deviations when the returns are correlated, i.e., when $\rho \neq 0$?

What about other allocations? For instance $(30\%, 70\%)$ or $(70\%, 30\%)$.

Solution. Let
X = return of small cap fund, $\mathbb{E}(X) = 0.12, \text{var}(X) = 0.16^2$
Y = return of large cap fund, $\mathbb{E}(Y) = 0.08, \text{var}(Y) = 0.11^2$

Return of (50%,50%) allocation:

$$0.5X + 0.5Y.$$

We have

$$\mathbb{E}(0.5X + 0.5Y) = 0.5\mathbb{E}(X) + 0.5\mathbb{E}(Y) = 0.5 \cdot 0.12 + 0.5 \cdot 0.08 = 0.10$$

and

$$\begin{aligned}\text{var}(0.5X + 0.5Y) &= 0.5^2\text{var}(X) + 0.5^2\text{var}(Y) + 2 \cdot 0.5 \cdot 0.5 \cdot \text{cov}(X, Y) \\ &= 0.25 \cdot 0.16^2 + 0.25 \cdot 0.11^2 = 0.009425\end{aligned}$$

So the standard deviation of return of the (50%,50%) allocation is $\sqrt{0.009425} = 0.097$.

Combinations of more than two random variables

Suppose X, Y, Z are uncorrelated with each other.

Then for all numbers a, b, c

$$\mathbb{E}(aX + bY + cZ) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y) + c \cdot \mathbb{E}(Z)$$

and

$$\mathsf{var}(aX + bY + cZ) = a^2 \cdot \mathsf{var}(X) + b^2 \cdot \mathsf{var}(Y) + c^2 \cdot \mathsf{var}(Z)$$

Similar formulas when we combine four or more uncorrelated variables.

When there is correlation

$$\begin{aligned}\mathsf{var}(aX + bY + cZ) = {} & a^2 \cdot \mathsf{var}(X) + b^2 \cdot \mathsf{var}(Y) + c^2 \cdot \mathsf{var}(Z) \\ & + 2ab \cdot \mathsf{cov}(X, Y) + 2bc \cdot \mathsf{cov}(Y, Z) + 2ac \cdot \mathsf{cov}(X, Z).\end{aligned}$$

Example

Suppose X_1, X_2, X_3 are uncorrelated to each other and all of them have mean μ and variance σ^2.

Then

$$\mathbb{E}\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{3}\left(\mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3)\right) = \mu$$

and

$$\mathsf{var}\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{3^2}\left(\mathsf{var}(X_1) + \mathsf{var}(X_2) + \mathsf{var}(X_3)\right) = \frac{\sigma^2}{3}.$$

Combinations of more random variables

Suppose $X_1, \ldots, X_n$ are random variables. Then for $a_1, \ldots, a_n \in \mathbb{R}$ we have

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \cdot \mathbb{E}(X_i)$$

and

$$\begin{aligned}\mathsf{var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \cdot a_j \cdot \mathsf{cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 \cdot \mathsf{var}(X_i) + 2 \cdot \sum_{1 \le i < j \le n} a_i \cdot a_j \cdot \mathsf{cov}(X_i, X_j)\end{aligned}$$

The last term vanishes if X_i and X_j are uncorrelated for all $i \neq j$.

Independence

Two random variables X, Y are *independent* if for all $A, B \subseteq \mathbb{R}$

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(x \in B).$$

Properties of independence

(a) Two random variables X, Y are independent if and only if for all $x, y \in \mathbb{R}$

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y).$$

(b) If X, Y are independent then

- $\text{cov}(X, Y) = 0$
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

Independence of discrete, continuous and normal variables

Discrete case

Suppose X, Y are discrete random variables. Then X, Y are independent if and only if for all $x, y \in \mathbb{R}$

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

Continuous case

Suppose X, Y are continuous random variables with joint density $f_{X,Y}$. Then X, Y are independent if and only if for all $x, y \in \mathbb{R}$

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

Normal variables

Suppose X, Y are normal and independent. Then $X + Y$ and $X - Y$ are normal.

Example (continued)

Suppose you can invest in the following two mutual funds:

- A *small cap equity* fund with annual expected return 12% and standard deviation 16%
- A *large cap equity* fund with annual expected return 8% and standard deviation 11%.

Suppose the above returns and normal and independent.

What is the probability that the annual return of the small cap equity fund is higher than that of the large cap equity fund?

Solution. Let X and Y denote the returns of the small and large cap funds respectively. We want $\mathbb{P}(X - Y > 0)$.

To that end, observe that $X - Y$ is normal with mean

$$\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = 0.12 - 0.08 = 0.04$$

and variance

$$\mathsf{var}(X - Y) = \mathsf{var}(X) + \mathsf{var}(Y) = 0.16^2 + 0.11^2 = 0.0377$$

Thus $X - Y \sim N(0.04, 0.0377)$ and

$$\begin{aligned} \mathbb{P}(X - Y > 0) &= 1 - \mathbb{P}(X - Y \leq 0) \\ &= 1 - \texttt{NORM.DIST}(\texttt{0}, \texttt{0.04}, \texttt{sqrt}(\texttt{0.0377}), \texttt{1}) \\ &= 0.5816 \end{aligned}$$

Independence of more than two random variables

The random variables $X_1, \ldots, X_n$ are independent if for all $A_1, \ldots, A_n \subseteq \mathbb{R}$

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n).$$

Sums and products of random variables

Suppose $X_1, \ldots, X_n$ are random variables. Then

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i).$$

If $X_1, \ldots, X_n$ are *independent* then we also have

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

Random sample (or i.i.d. random variables)

The random variables $X_1, \ldots, X_n$ are i.i.d. if

- They are independent
- They have the same distribution

If the distribution has cdf F we also write $X_1, X_2, \ldots, X_n \sim F$ and call $X_1, X_2, \ldots, X_n$ a **random sample of size n from F**.

Example of i.i.d. variables

Suppose there is a variable of interest X for each element of a large population.

Draw n random objects from the population (with replacement). Their values $X_1, \ldots, X_n$ are i.i.d.

Sample mean and sample variance

Suppose $X_1, X_2, \ldots, X_n$ are random variables.

The **sample mean** of $X_1, X_2, \ldots, X_n$ is

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Sometimes we write $\bar{X}_n$ to show the dependence on n.

The **sample variance** of $X_1, X_2, \ldots, X_n$ is

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}.$$

Expectation and variance of sample mean

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2 and

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then

$$\mathbb{E}(\bar{X}_n) = \mu$$

and

$$\mathsf{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Sample mean of i.i.d. normal variables

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. **normal** with mean μ and variance σ^2. Then

$$\bar{X}_n \sim N(\mu, \sigma^2/n).$$

Example

Suppose the adult men U.S. height distribution is normally distributed with a mean of 70 inches and standard deviation of 4 inches.

- We measure the height of a random adult male. What is the probability that the height is between 69 and 71 inches?
- We measure the heights of a random sample of 25 adult males. What is the probability that the sample mean is between 69 and 71 inches?
- We measure the heights of a random sample of 100 adult males. What is the probability that the sample mean is between 69 and 71 inches?

Solution. We have $\mu = 70, \sigma = 4$.

- Let X be the height of one randomly chosen adult male. Then $X \sim N(70, 4^2)$ and

$$\begin{aligned}\mathbb{P}(69 \leq X \leq 71) &= \mathbb{P}(X \leq 71) - \mathbb{P}(X \leq 69)\\ &= \texttt{NORM.DIST}(71, 70, 4, 1) - \texttt{NORM.DIST}(69, 70, 4, 1)\\ &= 0.197\end{aligned}$$

- For $n = 25$ we have $\bar{X}_n \sim N(70, 4^2/25)$. Thus

$$\begin{aligned}\mathbb{P}(69 \leq \bar{X}_n \leq 71) &= \mathbb{P}(\bar{X}_n \leq 71) - \mathbb{P}(\bar{X}_n \leq 69)\\ &= \texttt{NORM.DIST}(71, 70, 0.8, 1) - \texttt{NORM.DIST}(69, 70, 0.8, 1)\\ &= 0.7887\end{aligned}$$

- For $n = 100$ we have $\bar{X}_n \sim N(70, 4^2/100)$. Thus

$$\begin{aligned}\mathbb{P}(69 \leq \bar{X}_n \leq 71) &= \mathbb{P}(\bar{X}_n \leq 71) - \mathbb{P}(\bar{X}_n \leq 69)\\ &= \texttt{NORM.DIST}(71, 70, 0.4, 1) - \texttt{NORM.DIST}(69, 70, 0.4, 1)\\ &= 0.9876\end{aligned}$$

Sample mean of i.i.d. Bernoulli trials

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli with probability of success $p \in (0, 1)$.

Then $n\bar{X}_n \sim B(n, p)$. In particular,

$$\mathbb{E}(\bar{X}_n) = p \text{ and } \mathrm{var}(\bar{X}_n) = \frac{p(1-p)}{n}.$$

Suppose $p = 0.4$.

What is $\mathbb{P}(|\bar{X}_n - 0.4| \leq 0.01)$ for $n = 100, 1000, 10000$?

Law of Large Numbers and Central Limit Theorem

Law of Large Numbers

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ. Consider their sample mean

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Law of Large Numbers (informal)

As n grows, the distribution of $\bar{X}_n$ concentrates around μ.

Law of Large Numbers (precise)

As $n \to \infty$, the sample mean $\bar{X}_n$ **converges to μ in probability.** This means that for all $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Sample mean of i.i.d. normal variables

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. **normal** with mean μ and variance σ^2. Then

$$\bar{X}_n \sim N(\mu, \sigma^2/n).$$

Example

Suppose the adult men U.S. height distribution is normally distributed with a mean of 70 inches and standard deviation of 4 inches.

We measure the heights of a random sample of n adult males.

What is the probability that the sample mean is between 69 and 71 inches for $n = 100, 1000, 10000$?

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ, standard deviation σ, and **not** necessarily normal.

Recall the sample mean and its standardization

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}, \quad Z_n = \frac{\bar{X}_n - \mu_{\bar{X}_n}}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem (informal)

For large n the variable $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.

Central Limit Theorem (precise)

$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ **converges in distribution** to Z where $Z \sim N(0,1)$.

This means that as $n \to \infty$, the cdf of Z_n converges to the cdf of $N(0,1)$.

Central Limit Theorem (CLT)

Why should we care about the CLT?

- Sometimes we do not know a distribution but we can observe sample data $X_1, \ldots, X_n$.

  The CLT allows us to infer information about the unknown distribution from data.

- Sometimes an output depends on a complex combination of various inputs with known distributions.

  We can simulate (generate samples) of the output.

  The CLT allows us to draw conclusions from the simulation.

Convergence of random variables

Suppose $X_n,\ n = 1, 2, \ldots$ is a sequence of random variables and X is another random variable.

- We say that X_n converges to X in distribution, and write $X_n \rightsquigarrow X$, if
$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$
for all $x \in \mathbb{R}$ where F_X is continuous.

- We say that X_n converges to X in probability, and write $X_n \xrightarrow{\mathrm{P}} X$, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Fact: convergence in probability $\Rightarrow$ convergence in distribution.

Examples

- Suppose X, Y are two independent Bernoulli trials. Let $X_n = Y$ for $n = 1, 2, \ldots$. Then $X_n \rightsquigarrow X$ but

$$\mathbb{P}(|X_n - X| \geq 1) = \mathbb{P}(|Y - X| = 1) = 1/2.$$

  Thus $X_n \overset{\mathrm{P}}{\not\longrightarrow} X$.

- Suppose $X_n \sim N(0, 1/n)$. Then we have both

$$X_n \xrightarrow{\mathrm{P}} 0$$

  and

$$X_n \rightsquigarrow 0.$$

Proof of LLN

The proof of the LLN relies on the following inequalities, which are interesting on their own.

Markov's inequality

Suppose X is a non-negative random variable and $\mathbb{E}(X) < \infty$. Then for all $t > 0$

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

Chebyshev's inequality

Suppose X is a random variable with $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{var}(X)$ Then for all $t > 0$

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}.$$

Proof of LLN (finite variance)

Suppose $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance σ^2 and

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \text{ for } n = 1, 2, \ldots.$$

Suppose $\epsilon > 0$. From Chebyshev's inequality it follows that for all n

$$0 \leq \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since the right-hand-side tends to zero as $n \to \infty$ we conclude that

$$\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Technical note

With a more elaborate argument the assumption of finite variance can be dropped.

Intro to Prob and Stats, Week 4

Previously

- Multivariate joint distributions
- Independent random variables
- Random sample (or i.i.d. random variables)
- Law of Large Numbers (LLN), Central Limit Theorem (CLT)

This week

- Recap of i.i.d., LLN, CLT
- Moments and moment generating function
- If there is time: multivariate normal distribution
- Statistical inference

Recap of i.i.d., LLN, CLT

Random sample (or i.i.d. random variables)

The random variables $X_1, \ldots, X_n$ are independent identically distributed (i.i.d.) if

- They are independent
- They have the same distribution

If the distribution has cdf F we also write $X_1, X_2, \ldots, X_n \sim F$ and call $X_1, X_2, \ldots, X_n$ a **random sample of size n from F**.

Example of i.i.d. variables

Suppose there is a variable of interest X for each element of a large population.

Draw n random objects from the population (with replacement). Their values $X_1, \ldots, X_n$ are i.i.d.

Expectation and variance of sample mean

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2.

The **sample mean** of $X_1, X_2, \ldots, X_n$ is

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Observe:

$$\mathbb{E}(\bar{X}) = \mu$$

and

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Law of Large Numbers

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ. Consider their sample mean

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Law of Large Numbers (informal)

As n grows, the distribution of $\bar{X}_n$ concentrates around μ.

Law of Large Numbers (precise)

As $n \to \infty$, the sample mean $\bar{X}_n$ **converges to μ in probability.** This means that for all $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ, standard deviation σ, and **not** necessarily normal.

Recall the sample mean and its standardization

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}, \quad Z_n = \frac{\bar{X}_n - \mu_{\bar{X}_n}}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem (informal)

For large n the variable $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.

Central Limit Theorem (precise)

$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ **converges in distribution** to Z where $Z \sim N(0,1)$.

This means that as $n \to \infty$, the cdf of Z_n converges to the cdf of $N(0,1)$.

Convergence of random variables

Suppose $X_n,\ n = 1, 2, \ldots$ is a sequence of random variables and X is another random variable.

- We say that X_n converges to X in distribution, and write $X_n \rightsquigarrow X$, if
$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$
for all $x \in \mathbb{R}$ where F_X is continuous.

- We say that X_n converges to X in probability, and write $X_n \xrightarrow{\mathrm{P}} X$, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Fact: convergence in probability ⇒ convergence in distribution.

Examples

- Suppose X, Y are two independent Bernoulli trials. Let $X_n = Y$ for $n = 1, 2, \ldots$. Then $X_n \rightsquigarrow X$ but

  $$\mathbb{P}(|X_n - X| \geq 1) = \mathbb{P}(|Y - X| = 1) = 1/2.$$

  Thus $X_n \overset{\mathrm{P}}{\not\longrightarrow} X$.

- Suppose $X_n \sim N(0, 1/n)$. Then we have both

  $$X_n \xrightarrow{\mathrm{P}} 0$$

  and

  $$X_n \rightsquigarrow 0.$$

Proof of LLN

The proof of the LLN relies on the following inequalities, which are interesting on their own.

Markov's inequality

Suppose X is a non-negative random variable and $\mathbb{E}(X) < \infty$. Then for all $t > 0$

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

Chebyshev's inequality

Suppose X is a random variable with $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{var}(X)$ Then for all $t > 0$

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}.$$

Proof of LLN (finite variance)

Suppose $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance σ^2 and

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \quad \text{for} \quad n = 1, 2, \ldots .$$

Suppose $\epsilon > 0$. From Chebyshev's inequality it follows that for all n

$$0 \leq \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since the right-hand-side tends to zero as $n \to \infty$ we conclude that

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Technical note

With a more elaborate argument the assumption of finite variance can be dropped.

Moments and moment generating function

Moments of a random variable

Suppose X is a random variable.

Definition

For $k = 1, 2, \dots$ the k-th moment of X is $\mathbb{E}(X^k)$ provided $\mathbb{E}(|X|^k) < \infty$.

Observe

The first moment of X is $\mathbb{E}(X) = \mu_X$.
The second moment of X is $\mathbb{E}(X^2) = \sigma_X^2 + \mu_X^2$.

Definition

The k-th central moment of X is $\mathbb{E}((X - \mu_X)^k)$ provided $\mathbb{E}(|X - \mu_X|^k) < \infty$.

Skewness and kurtosis

Suppose X is a random variable with $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathsf{var}(X)$.

The *skewness* of X is the third standardized moment

$$\mathsf{skew}(X) := \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^3 .$$

The *kurtosis* of X is the fourth standardized moment

$$\mathsf{kurt}(X) := \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^4 .$$

Skewness is a measure of asymmetry of the distribution around μ.
Kurtosis is a measure of "tailedness" of the distribution.

Fact

If X is normal then $\mathsf{skew}(X) = 0$ and $\mathsf{kurt}(X) = 3$.

Moment generating function of a random variable

The moment generating function (mgf), or Laplace transform, of a random variable X is

$$\psi_X(t) = \mathbb{E}(e^{tX}).$$

Neat properties of the mgf

- ψ_X encodes the moments of X:

  $$\psi'_X(0) = \mathbb{E}(X),\ \psi''_X(0) = \mathbb{E}(X^2)$$

  and more generally $\psi_X^{(k)}(0) = \mathbb{E}(X^k)$ for $k = 1, 2, \ldots$
- If X_1, X_2 are independent random variables then

  $$\psi_{X_1+X_2}(t) = \psi_{X_1}(t) \cdot \psi_{X_2}(t).$$

- Suppose X, Y are random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in some open interval around 0 then $X \overset{d}{=} Y$.

Popular distributions and their mgf

Discrete

Distr.	pmf $\mathbb{P}(X = x)$	mgf $\psi(t)$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, \dots, n$	$(pe^t + (1-p))^n$
Geometric	$(1-p)^{x-1} p$ for $x = 1, 2, 3, \dots$	$\frac{pe^t}{1-(1-p)e^t}$ for $t < -\ln(1-p)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \dots$	$e^{\lambda(e^t - 1)}$

Continuous

Distribution	pdf $f(x)$	mgf $\psi(t)$
Uniform	$\frac{1}{b-a}$ for $x \in [a, b]$ and 0 ow	$\frac{e^{tb}-e^{ta}}{t(b-a)}$ for $t \neq 0$ and 1 ow
Exponential	$\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 ow	$\frac{\lambda}{\lambda - t}$ for $t < \lambda$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$e^{\mu t + \frac{\sigma^2 t^2}{2}}$

Sums of independent binomial/Poisson/normal

Suppose $X_i \sim B(n_i, p),\ i = 1, 2$ are independent. Then

$$\psi_{X_1}(t)\cdot\psi_{X_2}(t) = (pe^t+(1-p))^{n_1}\cdot(pe^t+(1-p))^{n_2} = (pe^t+(1-p))^{n_1+n_2}.$$

Thus $X_1 + X_2 \sim B(n_1 + n_2, p)$.

Suppose $X_i \sim \text{Pois}(\lambda_i),\ i = 1, 2$ are independent. Then

$$\psi_{X_1}(t) \cdot \psi_{X_2}(t) = e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}.$$

Thus $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Suppose $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1, 2$ are independent. Then

$$\psi_{X_1}(t) \cdot \psi_{X_2}(t) = e^{\mu_1 t+\frac{\sigma_1^2 t^2}{2}} \cdot e^{\mu_2 t+\frac{\sigma_2^2 t^2}{2}} = e^{(\mu_1+\mu_2)t+\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}.$$

Thus $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Convergence in distribution via mgf

Key technical lemma linked to convergence in distribution.

Lemma

Suppose $Z_1, Z_2, \ldots$ is a sequence of random variables and Z is another random variable.

If $\psi_{Z_n}(t) \to \psi(t)$ for t is some open interval around 0 then $Z_n \rightsquigarrow Z$.

Poisson as a limit of binomial

Suppose $\lambda > 0$ and $X_n \sim B(n, \lambda/n),\ n = 1, 2, \ldots$. Then as $n \to \infty$

$$\psi_{X_n}(t) = \left(\frac{\lambda}{n}e^t + 1 - \frac{\lambda}{n}\right)^n = \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \to e^{\lambda(e^t - 1)}$$

Thus $X_n \rightsquigarrow X$ for $X \sim \text{Pois}(\lambda)$.

Proof of CLT

Suppose $X_1, X_2, \dots, X_n$ are i.i.d. random variables with mean μ, standard deviation σ. Let ϕ be the mgf of $Y_i = (X_i - \mu)/\sigma$. Then

$$\phi'(0) = \mathbb{E}(Y_i) = 0,\ \phi''(0) = \mathbb{E}(Y_i^2) = 1,$$

and so

$$\begin{aligned}\phi(t) &= \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(0) + o(t^2)\\ &= 1 + \frac{t^2}{2} + o(t^2).\end{aligned}$$

For $n = 1, 2, \dots$ let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \text{ and } Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Proof of CLT (continued)

Key step

$Z_n = \dfrac{Y_1 + \cdots + Y_n}{\sqrt{n}}$ and thus $\psi_{Z_n}(t) = [\phi(t/\sqrt{n})]^n$.

Therefore

$$\begin{aligned} \psi_{Z_n}(t) &= [\phi(t/\sqrt{n})]^n \\ &= \left[1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \\ &\to e^{t^2/2}. \end{aligned}$$

In other words, ψ_n converges to the mgf of $N(0,1)$. Thus by the previous lemma $Z_n \rightsquigarrow Z$ for $Z \sim N(0,1)$.

If there is time: multivariate normal distribution

The bivariate normal distribution (simple case)

Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2$ and X_1, X_2 are independent.

Then X_1, X_2 have joint pdf

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}}$$

This joint pdf can be written in matrix-vector notation as follows

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \ \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

Some matrix algebra

Suppose $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \in \mathbb{R}^2,\ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$, $\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}$.

Then

$$\mathbf{a}^\mathsf{T}\mathbf{x} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = a_1x_1 + a_2x_2,$$

and

$$\mathbf{B}\mathbf{x} = \begin{bmatrix} b_{11}x_1 + b_{12}x_2 \\ b_{21}x_1 + b_{22}x_2 \end{bmatrix}.$$

In particular,

$$\mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x} = b_{11}x_1^2 + (b_{12} + b_{21})x_1x_2 + b_{22}x_2^2.$$

The bivariate normal distribution (general case)

The random variables X_1, X_2 have *joint bivariate normal distribution* if X_1, X_2 have joint density

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

for some

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \in \mathbb{R}^2 \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

with $\det(\Sigma) = \sigma_1^2 \cdot \sigma_2^2 - \sigma_{12}^2 > 0$.

Shorthand: $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\mathbf{X}$ is the random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Properties of the bivariate normal distribution

Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\mathbf{X}, \boldsymbol{\mu}, \Sigma$ are as above.

Then

(a) $\mathsf{cov}(X_1, X_2) = \sigma_{12}$

(b) X_1, X_2 are independent if and only if $\sigma_{12} = 0$

(c) for any $\mathbf{a} \in \mathbb{R}^2$ we have $\mathbf{a}^\mathsf{T}\mathbf{X} \sim N(\mathbf{a}^\mathsf{T}\boldsymbol{\mu}, \mathbf{a}^\mathsf{T}\Sigma\mathbf{a})$.

Multivariate normal distribution

The random variables $X_1, \ldots, X_n$ have *joint multivariate normal distribution* if $X_1, \ldots, X_n$ have joint density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ symmetric and positive definite.

Shorthand: $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\mathbf{X}$ is the random vector

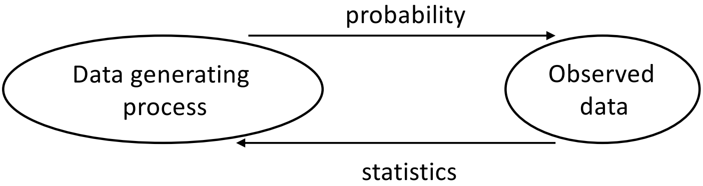
$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

Neat property

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{a} \in \mathbb{R}^n$ then $\mathbf{a}^\mathsf{T}\mathbf{X} \sim N(\mathbf{a}^\mathsf{T}\boldsymbol{\mu}, \mathbf{a}^\mathsf{T}\Sigma\mathbf{a})$.

Statistical inference

Recall our bird eye's view of probability and statistics



Probability (what we have done so far)

Formal mathematical framework to quantify uncertainty.

Statistical inference (what we will do next)

Use data to infer information about the underlying process that generated the data.

Statistical Inference

Goal

Make inference about an unknown probability distribution from observed data.

Random sample

Set $X_1, \ldots, X_n$ of i.i.d. random variables from some unknown distribution.

We often refer to $X_1, \ldots, X_n$ as **random sample** or as a **set of observations**.

Generic (and popular) example

Suppose there is a variable of interest X for each element of a large population.

Draw n random objects from the population. Their values $X_1, \ldots, X_n$ are i.i.d. with the same distribution as X.

In statistical inference

It is common to refer to the distribution generating the data as the *population distribution.*

Motivation

A candy makers produces candy bars advertised as containing 50 grams of candy. When the candy maker operation is working properly, the actual weight of a candy bar has a mean of 50 grams and a standard deviation of 2 grams.

For quality control, check samples of 25 candy bars periodically.

Questions

- What is the probability that the average weight of a sample of 25 candy bars is less than 49 grams?
- How far from 50 grams should the average weight of a 25-bar sample be before we raise a “red flag” about the candy maker operation?

Motivation (continued)

Example

An electric car maker just released its new Model X and advertises it as having excellent range per single charge.

Suppose we test a random sample of 16 Model X cars and obtain the following data (miles per single charge for each tested car):

```
234, 234, 247, 221, 247, 247, 325, 325, 312, 273,
286, 299, 286, 221, 286, 273
```

What is our best estimate of the average Model X miles of range in a single charge?

Can we estimate the standard deviation of miles of range as well?

Suppose the maker of Model X advertises it as having 295 miles of range in a single charge. Should we believe the advertisement?

Statistics of a random sample

Statistic

A number computed from a random sample.

Two important statistics

Suppose $X_1, \ldots, X_n$ is a random sample.

- Sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

- Sample variance

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}.$$

Sample mean

Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2.

The **sample mean** of these random variables is

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n}.$$

Recall

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

Implication of the Central Limit Theorem

If the population has mean μ and variance σ^2 then CLT implies

$$\bar{X} \approx N(\mu, \sigma^2/n) \iff \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1).$$

Example

A candy manufacturer makes candy bars. When the packing operation is properly working, the mean weight of a candy bar is $\mu = 50$ grams with standard deviation $\sigma = 2$ grams.

For quality control, take periodic samples of 25 candy bars and compute the average weight $\bar{X}$ of the 25 candy bars.

Find the probability of each of the following events (assume CLT applies):

- The sample mean $\bar{X}$ is less than 49 grams
- The sample mean $\bar{X}$ is between 49.6 grams and 50.4 grams.

Find a symmetric interval around $\mu = 50$ grams that includes the sample mean $\bar{X}$ with probability 0.95.

Solution.

We have $\mu = 50, \sigma = 2$ and thus by the CLT

$$\bar{X} \approx N(50, 2^2/25) = N(50, 4/25).$$

Hence

$$\mathbb{P}(\bar{X} \leq 49) = \texttt{NORM.DIST}(49, 50, 2/5, 1) = 0.0062$$

and

$$\begin{aligned} &\mathbb{P}(49.6 \leq \bar{X} \leq 50.4) \\ &= \texttt{NORM.DIST}(50.4, 50, 2/5, 1) - \texttt{NORM.DIST}(49.6, 50, 2/5, 1) \\ &= 0.68269. \end{aligned}$$

Next we want a such that

$$\mathbb{P}(\mu - a \leq \bar{X} \leq \mu + a) = 0.95$$

Thus $a = 1.96 \cdot 2/5 = 0.784$. The sample mean is in the interval $[49.216, 50.784]$ with probability 0.95.

Acceptance intervals (used in quality control)

Given $\alpha \in (0,1)$, let $z_{\alpha/2} := (1-\alpha/2)$-quantile of standard normal:

$$z_{\alpha/2} := \texttt{NORM.S.INV}(1-\alpha/2) = \texttt{qnorm}(1-\alpha/2).$$

Then by CLT

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

We conclude that with probability $1-\alpha$

$$\mu - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

To state the above, we say

$$\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\mu - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

is a $(1-\alpha)$ acceptance interval.

Acceptance intervals

Suppose a population has mean μ and variance σ^2 and $X_1, X_2, \ldots, X_n$ is a random sample from the population.

Because of the CLT

With probability $1 - \alpha$ the sample mean $\bar{X}$ is in the interval

$$\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\mu - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

for $z_{\alpha/2} = \texttt{NORM.S.INV}(1 - \alpha/2) = \texttt{qnorm}(1 - \alpha/2)$.

In the previous candy manufacturer example.

We have $\mu = 50, \sigma = 2$, and $n = 25$.

- 0.95 acceptance interval: want $(1-\alpha)$ acceptance interval for $\alpha = 0.05$. First compute

  $z_{0.05/2} = \texttt{NORM.S.INV}(1{-}0.05/2) = \texttt{NORM.S.INV}(975) = 1.96.$

  0.95 acceptance interval:

  $$50 \pm 1.96 \cdot \frac{2}{5} = [49.216, 50.784].$$

- 0.99 acceptance interval: want $(1-\alpha)$ acceptance interval for $\alpha = 0.01$. First compute

  $z_{0.01/2} = \texttt{NORM.S.INV}(1{-}0.01/2) = \texttt{NORM.S.INV}(995) = 2.576.$

  0.99 acceptance interval:

  $$50 \pm 2.576 \cdot \frac{2}{5} = [48.9696, 51.0304].$$

Main types of statistical inference

Point estimation

Provide a “best guess” of some parameter θ of the distribution. For instance, the mean or variance of the distribution.

Confidence interval

Construct an interval that traps a parameter θ with some (typically high) probability.

Hypothesis testing

Start with a default theory (a “null hypothesis”) about the distribution. Then test if the data provides sufficient evidence to reject the theory.

Point estimation

Suppose $X_1, \ldots, X_n$ is a random sample from some distribution F and supposed θ is some parameter of F.

A *point estimator* $\widehat{\theta}_n$ of θ is some function of $X_1, \ldots, X_n$

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

Sometimes we drop the subindex and write $\widehat{\theta}$ for $\widehat{\theta}_n$.

Examples of point estimators

- The sample mean is a point estimator of the mean μ:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

- The sample variance is a point estimator of the variance σ^2:

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1}.$$

Sampling distribution

Suppose $\widehat{\theta}_n$ is a point estimator of a parameter θ. The parameter θ is fixed but unknown while $\widehat{\theta}_n$ is random. The distribution of $\widehat{\theta}_n$ is the *sampling distribution*.

The **bias** of $\widehat{\theta}_n$ is

$$\mathsf{bias}(\widehat{\theta}_n) = \mathbb{E}(\widehat{\theta}_n) - \theta.$$

The **standard error** of $\widehat{\theta}_n$ is its standard deviation:

$$\mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathsf{var}(\widehat{\theta}_n)}$$

Distribution of the sample mean and sample variance

Suppose $X_1, \ldots, X_n$ is a random sample from a distribution with mean μ and variance σ^2.

Then the sample mean $\bar{X}$ has mean μ and variance σ^2/n, that is:

$$\text{bias}(\bar{X}) = 0 \text{ and } \text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Also the sample variance has mean σ^2, that is:

$$\text{bias}(S^2) = 0.$$

Sampling from a normal distribution

Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$.

Then the sample mean $\bar{X}$ and sample variance S^2 are independent. Furthermore, $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2$ has chi-square distribution with $n-1$ degrees of freedom.

The chi-square distribution

A continuous random variable has chi-square distribution with m degrees of freedom if it has the following density

$$f(x) = \begin{cases} \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

where $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is the function $\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$.
This function satisfies $\Gamma(x+1) = x\Gamma(x)$ for all $x > 0$.
In particular $\Gamma(k+1) = k!$ for $k = 1, 2, \ldots$.

Connection between normal and chi-square

If $Y_1, \ldots, Y_m$ are iid standard normal then $Y_1^2 + \cdots + Y_m^2$ has chi-square distribution with m degrees of freedom.

Properties of the chi-square distribution

Suppose X has chi-square distribution with m degrees of freedom. Then

$$\mathbb{E}(X) = m \text{ and } \operatorname{var}(X) = 2m.$$

Moment generating function:

$$\psi_X(t) = (1 - 2t)^{-m/2} \text{ for } t < 1/2.$$

R functions for density, cdf, quantile, random draw:

$\texttt{dchisq}(\texttt{x}, \texttt{df})$, $\texttt{pchisq}(\texttt{x}, \texttt{df})$, $\texttt{qchisq}(\texttt{p}, \texttt{df})$, $\texttt{rchisq}(\texttt{n}, \texttt{df})$

Unbiasedness, consistency, mean squared error

Suppose $\widehat{\theta}_n$ is a point estimator of a parameter θ.

- The estimator $\widehat{\theta}_n$ is **unbiased** if bias$(\widehat{\theta}_n) = 0$, i.e., $\mathbb{E}(\widehat{\theta}_n) = \theta$.
- The estimator $\widehat{\theta}_n$ is **consistent** if $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.
- The **mean squared error** of $\widehat{\theta}_n$ is

$$\mathsf{MSE}(\widehat{\theta}_n) := \mathbb{E}((\widehat{\theta}_n - \theta)^2).$$

Bias-variance decomposition

The mean-squared error satisfies

$$\mathsf{MSE}(\widehat{\theta}_n) = \mathsf{var}(\widehat{\theta}_n) + (\mathsf{bias}(\widehat{\theta}_n))^2$$

Confidence intervals for the mean

The sample mean $\bar{X}$ is an **estimator** of μ: each realization of $\bar{X}$ gives a **point estimate** of μ.

Acceptance and confidence intervals account for the estimator's variability.

Assume population has mean μ and variance σ^2.

Acceptance interval: interval around the mean μ

$$\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

With probability $1 - \alpha$ this interval brackets $\bar{X}$.

Confidence interval: interval around the sample mean $\bar{X}$

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

With probability $1 - \alpha$ this interval brackets μ.

Confidence interval terminology

Standard error of sample mean: $\text{se}(\bar{X}) = \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$.

Recall $(1-\alpha)$ confidence interval:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

Lower Confidence Limit (LCL): $\bar{X} - z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

Upper Confidence Limit (UCL): $\bar{X} + z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

Margin of error at the $(1-\alpha)$ conf level: $z_{\alpha/2} \cdot \text{se}(\bar{X}) = z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

Example

Suppose the standard deviation of the mileage per single charge of Model X cars is 30 miles.

The sample mean of the mileage per single charge of 16 randomly chosen Model X cars is 269.75.

Form a 95% confidence interval for the population mean of the mileage per single charge of Model X cars.

Solution. The 95% confidence interval is $\bar{X} \pm z_{0.025} \cdot \sigma/\sqrt{n}$ that is

$$269.75 \pm 1.96 \cdot \frac{30}{\sqrt{16}} = 269.75 \pm 14.7 = [245.05, 284.45]$$

Wrinkle

When analyzing data, we usually do not know the population variance σ^2. Need to estimate it.

Confidence interval for population mean

Suppose $X_1, \ldots, X_n$ is a random sample from a population with unknown mean μ and unknown variance σ^2.

Use **sample variance** $S^2 = \dfrac{1}{n-1} \cdot \displaystyle\sum_{i=1}^{n} (X_i - \bar{X})^2$ to estimate σ^2.

Fact: S^2 is a random variable with $\mathbb{E}(S^2) = \sigma^2$.

Version 1: quick and dirty, ok for large samples

- Compute sample mean $\bar{X} = \dfrac{X_1 + \cdots + X_n}{n}$
- Compute sample variance $S^2 = \dfrac{1}{n-1} \cdot \displaystyle\sum_{i=1}^{n} (X_i - \bar{X})^2$
- $(1-\alpha)$ confidence interval for μ:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

Excel command for sample variance: `VAR.S(...)`

R command for sample variance: `var(...)`

Confidence interval for population mean

Version 2: refined for small samples

- Compute sample mean $\bar{X} = \dfrac{X_1 + \cdots + X_n}{n}$
- Compute sample variance $S^2 = \dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$
- $(1-\alpha)$ confidence interval for μ:

$$\bar{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

- t_{n-1}: t-distribution with $n-1$ degrees of freedom
- $t_{n-1,\alpha/2} = (1-\alpha/2)$-quantile of t_{n-1}.

  Excel command: $t_{n-1,\alpha/2} = \mathtt{T.INV}(\mathtt{1} - \alpha/\mathtt{2}, \mathtt{n} - \mathtt{1})$

  R command: $t_{n-1,\alpha/2} = \mathtt{qt}(\mathtt{1} - \alpha/\mathtt{2}, \mathtt{n} - \mathtt{1})$

Normal distribution and t-distribution

Assume the population is normal with mean μ and variance σ^2.

Then for a random sample $X_1, \ldots, X_n$ the Z-statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is standard normal.

On the other hand, the t-statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is not standard normal because there is some extra variability in S.

The t-statistic has a t-distribution with $n - 1$ degrees of freedom.

The t-distribution

Let $\nu > 0$. A continuous random variable X has t-distribution with ν degrees of freedom if it has density

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} .$$

Shorthand: $X \sim t_\nu$.

Expectation

$$\mathbb{E}(X) = \begin{cases} 0 & \text{for } \nu > 1 \\ \text{undefined} & \text{for } 0 < \nu \leq 1 \end{cases}$$

Variance

$$\text{var}(X) = \begin{cases} \frac{\nu}{\nu-2} & \text{for } \nu > 2 \\ \infty & \text{for } 1 < \nu \leq 2 \\ \text{undefined} & \text{for } 0 < \nu \leq 1 \end{cases}$$

The moment generating function is not defined for $t \neq 0$.

Special cases of the t-distribution

As $\nu \to \infty$ the t-distribution approaches the standard normal distribution:

$$\lim_{\nu\to\infty} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

When $\nu = 1$ we get the *Cauchy* distribution

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

R functions for density, cdf, quantile, random draw:

$$\texttt{dt}(\texttt{x},\texttt{df}),\ \texttt{pt}(\texttt{x},\texttt{df}),\ \texttt{qt}(\texttt{p},\texttt{df}),\ \texttt{rt}(\texttt{n},\texttt{df})$$

Excel functions for density, cdf, quantile:

$$\texttt{T.DIST}(\texttt{x},\texttt{df},\texttt{0}),\ \texttt{T.DIST}(\texttt{x},\texttt{df},\texttt{1}),\ \texttt{T.INV}(\texttt{x},\texttt{df})$$

Connection with normal and chi-square

Suppose Y, Z are independent, Y has χ^2 distribution with m degrees of freedom, and Z is standard normal. Then

$$\frac{Z}{\sqrt{Y/m}}$$

has t-distribution with m degrees of freedom.

In particular, if $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then the t-statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

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

Critical values $z_{\alpha/2}$ and $t_{\nu,\alpha/2}$

Suppose $\alpha \in (0,1)$. The expressions $z_{\alpha/2}$ and $t_{\nu,\alpha/2}$ denote the following quantiles of the standard norm and t-distributions.

$z_{\alpha/2} = (1-\alpha/2)$-quantiles of standard normal

$$z_{\alpha/2} = \texttt{NORM.S.INV}(1-\alpha/2) = \texttt{qnorm}(1-\alpha/2)$$

$t_{\nu,\alpha/2} = (1-\alpha/2)$-quantiles of t-distribution with ν d.f.

$$t_{\nu,\alpha/2} = \texttt{T.INV}(1-\alpha/2,\nu) = \texttt{qt}(1-\alpha/2,\nu)$$

Similar commands for the cdf. If $Z \sim N(0,1)$ and $T \sim t_\nu$ then

$$\mathbb{P}(Z \le x) = \texttt{NORM.S.DIST}(x,1) = \texttt{pnorm}(x)$$

and

$$\mathbb{P}(T \le x) = \texttt{T.DIST}(x,\nu,1) = \texttt{pt}(x).$$

Example

Suppose we measure the mileage per single charge of a random sample of 16 Model X cars.

Suppose the sample mean and sample standard deviation are respectively 269.75 and 35.0438, and the random sample included 16 observations.

Construct 95% and 99% confidence intervals for the population mean Mileage μ.

Solution

0.95 confidence level:

$$269.75 \pm t_{15,0.025} \cdot \frac{35.0438}{\sqrt{16}} = [251.0765, 288.4235]$$

0.99 confidence level:

$$269.75 \pm t_{15,0.005} \cdot \frac{35.0438}{\sqrt{16}} = [243.934, 295.566]$$

Intro to Prob and Stats, Week 5

Last week

- Joint probability distributions
- Central Limit Theorem
- Statistical inference

This week

- Confidence intervals
- Hypothesis testing
- Regression

Confidence intervals

Motivation

- An electric car maker just released its new Model X and advertises it as having excellent range per single charge.

  To estimate the average range per single charge, we test a random sample of 16 Model X cars.

- A political campaign needs to assess the popularity of their candidate in a swing state (e.g., Pennsylvania). To that end, they survey a random set of potential voters.

Recall main types of statistical inference

Goal of statistical inference: use data to make inferences about an unknown “population” distribution.

Point estimation

Provide a “best guess” of some parameter of the distribution. For instance, the mean or variance of the distribution.

Confidence interval

Construct an interval that traps a parameter with some (typically high) probability.

Hypothesis testing

Start with a default theory (a “null hypothesis”) about the distribution. Test if the data provides sufficient evidence to reject the theory.

Two important point estimators

Suppose $X_1, \ldots, X_n$ is a random sample from some distribution.

- The sample mean is an unbiased point estimator of the mean μ:
$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$
- The sample variance is an unbiased point estimator of the variance σ^2:
$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}.$$

Excel commmands

Sample mean: `AVERAGE(...)`, sample variance: `VAR.S(...)`

R commmands

Sample mean: `mean(...)`, sample variance: `var(...)`

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, and **not** necessarily normal.

Recall the sample mean and its standardization

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}, \quad Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem (informal)

For large n the variable $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$.

Central Limit Theorem (precise)

As $n \to \infty$, the cdf of $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to the cdf of $N(0,1)$.

Confidence intervals

Suppose a population has mean μ and variance σ^2 and $X_1, X_2, \dots, X_n$ is a random sample from the population.

Because of the CLT

With probability $(1-\alpha)$ the following interval traps μ:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

for $z_{\alpha/2} = \texttt{NORM.S.INV}(1-\alpha/2) = \texttt{qnorm}(1-\alpha/2)$.

The above interval comes from

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \Rightarrow \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Example

Suppose the standard deviation of the mileage per single charge of Model X cars is 30 miles.

The sample mean of the mileage per single charge of 16 randomly chosen Model X cars is 269.75.

Form a 95% confidence interval for the population mean of the mileage per single charge of Model X cars.

Solution. The 95% confidence interval is $\bar{X} \pm z_{0.025} \cdot \sigma/\sqrt{n}$ that is

$$269.75 \pm 1.96 \cdot \frac{30}{\sqrt{16}} = 269.75 \pm 14.7 = [245.05, 284.45]$$

Wrinkle

When analyzing data, we usually do not know the population variance σ^2. Need to estimate it.

Confidence interval for population mean

Suppose $X_1, \ldots, X_n$ is a random sample from a population with unknown mean μ and unknown variance σ^2.

Use **sample variance** $S^2 = \dfrac{1}{n-1} \cdot \displaystyle\sum_{i=1}^{n} (X_i - \bar{X})^2$ to estimate σ^2.

When we know σ

Get confidence interval from the Z-statistic and CLT

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

When we do not know σ

Use instead the T-statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \approx t_{n-1}.$$

Confidence intervals for the population mean

When we know σ

Use $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$ to get $(1-\alpha)$ confidence interval:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

When we do not know σ

Use $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \approx t_{n-1}$ to get $(1-\alpha)$ confidence interval:

$$\bar{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

Notation:

- t_{n-1}: t-distribution with $n-1$ degrees of freedom
- $t_{n-1,\alpha/2} = (1-\alpha/2)$-quantile of t_{n-1}.
- Use ν or df to denote the degrees of freedom.

Critical values $z_{\alpha/2}$ and $t_{\mathsf{df},\alpha/2}$

Suppose $\alpha \in (0,1)$. The expressions $z_{\alpha/2}$ and $t_{\mathsf{df},\alpha/2}$ denote the following quantiles of the standard normal and t-distributions.

$z_{\alpha/2} = (1-\alpha/2)$-quantile of standard normal

$$z_{\alpha/2} = \texttt{NORM.S.INV}(1-\alpha/2) = \texttt{qnorm}(1-\alpha/2)$$

$t_{\mathsf{df},\alpha/2} = (1-\alpha/2)$-quantile of t-distribution

$$t_{\mathsf{df},\alpha/2} = \texttt{T.INV}(1-\alpha/2, \mathsf{df}) = \texttt{qt}(1-\alpha/2, \mathsf{df})$$

Similar commands for the cdf. If $Z \sim N(0,1)$ and $T \sim t_{\mathsf{df}}$ then

$$\mathbb{P}(Z \leq x) = \texttt{NORM.S.DIST}(x, 1) = \texttt{pnorm}(x)$$

and

$$\mathbb{P}(T \leq x) = \texttt{T.DIST}(x, \mathsf{df}, 1) = \texttt{pt}(x, \mathsf{df}).$$

Example

Suppose we measure the mileage per single charge of a random sample of 16 Model X cars.

Suppose the sample mean and sample standard deviation are respectively 269.75 and 35.0438, and the random sample included 16 observations.

Construct 95% and 99% confidence intervals for the population mean Mileage μ.

Solution

0.95 confidence level:

$$269.75 \pm t_{15,0.025} \cdot \frac{35.0438}{\sqrt{16}} = [251.0765, 288.4235]$$

0.99 confidence level:

$$269.75 \pm t_{15,0.005} \cdot \frac{35.0438}{\sqrt{16}} = [243.934, 295.566]$$

Population proportion

For a proportion p of the population we have $X = 1$.
For the remaining proportion $1 - p$ we have $X = 0$.

In other words, for a randomly chosen object we have

$$X = \begin{cases} 1 & \text{with probability } \; p \\ 0 & \text{with probability } \; 1 - p. \end{cases}$$

In this case $\mathbb{E}(X) = p$ and $\mathrm{var}(X) = p(1 - p)$.

Sample proportion

Given a random sample $X_1, \dots, X_n$ from X, the **sample proportion** is

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n} = \frac{\text{number of successes in the sample}}{n}.$$

Confidence interval for population proportion

For estimating proportion, CLT implies that $\hat{p} \approx N(p, p(1-p)/n)$.

Thus proceed as follows:

- Compute sample proportion $\hat{p}$.
- Use $\hat{p}(1-\hat{p})$ to estimate population variance $p(1-p)$.
- $(1-\alpha)$ confidence interval for p:

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

- The **margin of error** at the $(1-\alpha)$ confidence level is

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

Example

Suppose in a poll of 100 randomly selected Americans, 41 of them approve of Biden's job.

Construct a 95% confidence interval for the proportion of Americans who approve of Biden's job.

Solution.
0.95 confidence interval:

$$0.41 \pm z_{0.025} \cdot \frac{\sqrt{0.59 \cdot 0.41}}{\sqrt{100}} = 0.41 \pm 1.96 \cdot 0.04918 = [0.3136, 0.5064]$$

MOE at the 0.95 confidence level is $1.96 \cdot 0.04918 = 0.0964$

Question

Find the smallest sample size n that guarantees a margin of error of 0.04 or less at the 0.95 confidence level.

Solution. We want MOE $= z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.04$. Therefore we want

$$\sqrt{n} \geq \frac{z_{0.025} \cdot \sqrt{\hat{p}(1-\hat{p})}}{0.04} = \frac{1.96 \cdot \sqrt{\hat{p}(1-\hat{p})}}{0.04}$$

The max value of $\hat{p}(1-\hat{p})$ is 0.25 when $\hat{p} = 0.5$.

Thus we need

$$\sqrt{n} \geq \frac{1.96 \cdot \sqrt{0.25}}{0.04} = 24.5$$

and so $n \geq 600.25$. Hence the smallest sample size is 601.

Selecting the sample size

Suppose we want a margin of error to have at most a desired level D. How should we choose the sample size n?

If we know the population standard deviation

Recall the margin of error at the $(1-\alpha)$ confidence level:

$$z_{\alpha/2} \cdot \mathsf{se}(\bar{X}) = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

We want this quantity to be at most D so we need to choose

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{D}\right)^2 .$$

When we do not know σ, use some reasonable estimate $\widehat{\sigma}$ instead.

Special case: population proportion

In this case we can choose $\widehat{\sigma} = \sqrt{0.25} = 0.5$.

Hypothesis Testing

Hypothesis testing

Hypothesis

Statement about a population.

Hypothesis test

A method to determine the evidence in data related to a hypothesis.

Formulating a hypothesis test

- Specify a **null hypothesis:** a hypothesized value for a population parameter.
- Choose the probability $\alpha \in (0, 1)$ of **rejecting** the null hypothesis when it is true.
  - We want this probability to be small, e.g., $\alpha = 0.05$.
  - The probability α is called the **significance level**

Implementing a hypothesis test

- Analogy to a trial: assume the null hypothesis unless find sufficient evidence against it.
- For a null hypothesis on a population parameter, and a significance level α, we have three equivalent testing methods

  Reject the null hypothesis if:

  - The $(1-\alpha)$ confidence interval does not include the null value.
  - The absolute value of the t-statistic exceeds the critical t-value
  - The p-value is less than α

Next, we will discuss testing a hypothesis on the population mean. That is, a null hypothesis of the form

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0.$$

Hypothesis testing

To test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the α significance level:

- Confidence interval: Reject if

$$\mu_0 \notin \left[\bar{X} - t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

- t-statistic: Compute t-stat $= \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ and reject if

$$|\text{t-stat}| > t_{n-1,\alpha/2}$$

- p-value: Compute p-value $= \mathbb{P}(|t_{n-1}| > |\text{t-stat}|)$ and reject if

$$\text{p-value} < \alpha.$$

Example

The maker of the all-electric Model X cars claim that their cars' average range per full charge is 295 miles.

Null hypothesis $H_0 : \mu = 295.$

Alternative hypothesis $H_1 : \mu \neq 295.$

Suppose the sample mean and sample standard deviation are respectively 269.75 and 35.0438, and the random sample included 16 observations.

Do we reject the null hypothesis at the $\alpha = 5\%$ significance level?

Do we reject the null hypothesis at the $\alpha = 1\%$ significance level?

First method: via confidence interval

Reject the null hypothesis if $(1-\alpha)$ confidence interval does not contain the null value μ_0.

Mileage example

0.95 confidence level:
$269.75 \pm t_{15,0.025} \cdot \frac{35.0438}{\sqrt{16}} = [251.0765, 288.4235]$

Reject the null hypothesis $H_0 : \mu = 295$ at the 5% significance level.

0.99 confidence level:
$269.75 \pm t_{15,0.005} \cdot \frac{35.0438}{\sqrt{16}} = [243.934, 295.566]$

Do not reject the null hypothesis $H_0 : \mu = 295$ at the 1% significance level.

Second method: via the t-statistic

Reject the null hypothesis if the absolute value of the t-statistic

$$\text{t-stat} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

exceeds the critical value $t_{n-1,\alpha/2}$.

Mileage example

T-statistic

$$\text{t-stat} = \frac{269.75 - 295}{35.0438/\sqrt{16}} = -2.8821$$

Since $|\text{t-stat}| = 2.8821 > 2.1314495 = t_{15,0.025}$, we reject the null hypothesis $H_0 : \mu = 295$ at the 5% significance level.

Since $|\text{t-stat}| = 2.8821 < 2.946712883 = t_{15,0.005}$, we do not reject the null hypothesis $H_0 : \mu = 295$ at the 1% significance level.

Third method: via the p-value

Reject the null hypothesis if the

$$\text{p-value} = \mathbb{P}(|t_{n-1}| > |\text{t-stat}|)$$

is less than the significance level α.

Mileage example

$$\text{p-value} = \mathbb{P}(|t_{15}| > 2.8821) = 0.0114$$

We thus reject at the 5% significance level but do not reject at the 1% significance level.

To compute the p-value

Excel command: $\text{p-value} = 2 * (1 - \texttt{T.DIST}(|\text{t-stat}|, \texttt{n} - 1, 1))$

R command: $\text{p-value} = 2 * (1 - \texttt{pt}(|\text{t-stat}|, \texttt{n} - 1))$

Hypothesis testing when population variance is known

When the population variance σ^2 is known, testing a hypothesis is simpler.

To test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the α significance level:

- Confidence interval: Reject if

$$\mu_0 \notin \left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

- z-statistic: Compute z-stat $= \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ and reject if

$$|\text{z-stat}| > z_{\alpha/2}$$

- p-value: Compute p-value $= \mathbb{P}(|Z| > |\text{z-stat}|)$ for $Z \sim N(0, 1)$ and reject if

$$\text{p-value} < \alpha.$$

Hypothesis testing for population proportion

To test $H_0 : p = p_0$ against $H_1 : p \neq p_0$ at the α significance level proceed assuming the variance is known to be $p_0(1 - p_0)$:

- Confidence interval: Reject if

$$p_0 \notin \left[\hat{p} - z_{\alpha/2} \cdot \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \cdot \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right]$$

- z-statistic: Compute z-stat $= \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$ and reject if

$$|\text{z-stat}| > z_{\alpha/2}$$

- p-value: Compute p-value $= \mathbb{P}(|Z| > |\text{z-stat}|)$ for $Z \sim N(0, 1)$ and reject if

$$\text{p-value} < \alpha.$$

Example

An experienced gambler claims to have a strategy for blackjack. He claims that his strategy allows him to win 60% of the time.

Suppose you observe the gambler play 200 blackjack hands and he wins 105 of them. Should you believe his claim?

Test $H_0 : p = 0.6$ against $H_1 : p \neq 0.6$ at the 0.05 significance level.

Example (continued)

0.95 confidence interval:

$$0.525 \pm z_{0.025} \cdot \frac{\sqrt{0.6 \cdot 0.4}}{\sqrt{200}} = [0.4571, 0.5929] \Rightarrow \text{ reject}$$

z-statistic:

$$\text{z-stat} = \frac{0.525 - 0.6}{\frac{\sqrt{0.6 \cdot 0.4}}{\sqrt{200}}} = -2.165$$

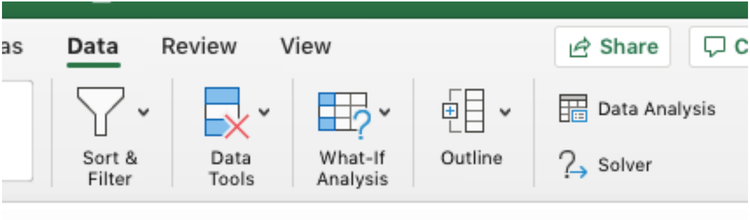
Since $|\text{z-stat}| = 2.165 > 1.96 = z_{0.025}$, we reject.

p-value:

$$\text{p-value} = \mathbb{P}(|Z| > 2.165) = 0.03038 < 0.05 \Rightarrow \text{reject}$$

Data analysis in Excel

Sometimes it is convenient to do data analysis in Excel. To that end, install the `Data Analysis ToolPak`. When you are done, under the "Data" tab your Excel should look something like this:



Related R and Excel commands

The previous procedure for hypothesis testing based on confidence intervals, t-stat, and p-value is called a "t-test".

R command for t-test

```
t.test(...)
```

Related functionality in Excel

```
Data -> Data Analysis -> Descriptive Statistics
```

One-sided hypothesis tests

The previous type of tests are “two-sided” hypothesis tests. These are the most common.

Sometimes, we are interested in testing a null hypothesis

$$H_0 : \mu \leq \mu_0$$

against the alternative

$$H_1 : \mu > \mu_0.$$

Proceed as when testing $H_0 : \mu = \mu_0$ but reject only on one side.

One-sided hypothesis tests

To test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ at the α significance level:

- Compute t-stat $= \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Reject if t-stat $> t_{n-1,\alpha}$

To test $H_0 : \mu \geq \mu_0$ against $H_1 : \mu < \mu_0$

- Compute t-stat $= \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Reject if t-stat $< -t_{n-1,\alpha}$

Similar procedures to test $H_0 : p \geq p_0$ or $H_0 : p \leq p_0$ for a population proportion p.

Mileage Example

Test the null hypothesis $H_0 : \mu \geq 295$ against $H_1 : \mu < 295$.

We have t-stat$= -2.8821 < -1.7305 = -t_{15,0.05}$.
Thus reject at the 5% significance level.

Gambler's Example

Out of 200 blackjack hands, gambler wins 105 of them.
Test $H_0 : p \geq 0.6$ against $H_1 : p \leq 0.6$ at the 0.05 significance level.

We have z-stat$= -2.165 < -1.644 = -z_{0.05}$.
Thus reject at the 5% significance level.

Possible outcomes of a hypothesis test

	H_0 is true	H_0 is false
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Significance level α of a test

This is the probability of rejecting a true hypothesis, or equivalently the probability of making a Type I error.

Power of a test

The probability of rejecting a false hypothesis.

We generally want high power and low significance but there is a tradeoff between these two goals.

Inference about joint distributions

Goal: make inference about the joint distribution of two random variables X, Y from a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Motivating example

An individual can invest capital in a variety of mutual funds. For diversification purposes, it is attractive to identify mutual funds with returns that have little or no correlation.

- Consider the historical data of monthly returns of some funds from the Vanguard family. (See the data in `VanguardFunds`.)
- How can we use that data to infer what returns are correlated/uncorrelated?

Another motivating example

A consumer credit reporting agency finds that American consumers make average monthly debt payments of \$983. However, the amount a consumer pays depends a great deal on where the consumer lives. (See the data in `DebtPayments`.)

Variance, covariance, correlation revisited

Suppose X, Y are random variables and

$$\mu_X = \mathbb{E}(X),\ \mu_Y = \mathbb{E}(Y).$$

Recall

The **variance** of X (often denoted σ_X^2) is

$$\mathsf{var}(X) = \mathbb{E}((X - \mu_X)^2).$$

Likewise for Y.

The **covariance** of X, Y (often denoted $\sigma_{X,Y}$) is

$$\mathsf{cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

The **correlation** of X, Y is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}.$$

The correlation is always between -1 and 1.

Estimators of covariance and correlation

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from the joint distribution of two random variables X, Y.

The **sample covariance** of X, Y is

$$S_{XY} := \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}.$$

The **sample correlation** is

$$r_{XY} := \frac{S_{XY}}{S_X \cdot S_Y}.$$

Excel commands: `COVARIANCE.S(...)` and `CORREL(...)`

R commands: `cov(...)` and `cor(...)`

Testing hypotheses about correlation

- Common null hypothesis

$$H_0 : \rho_{XY} = 0$$

- If X and Y are normally distributed, or if n is large, we can test H_0 using:

$$\text{t-stat} := r_{XY} \cdot \sqrt{\frac{n-2}{1-r_{XY}^2}}$$

  Under the null hypothesis this statistic has t-distribution with $n-2$ degrees of freedom.

- For a two-tailed test, reject H_0 if the p-value $< \alpha$, or equivalently, if $|\text{t-stat}| > t_{n-2,\alpha/2}$.

Conditional distributions

Suppose X, Y are random variables. We want to make sense of the conditional distribution of one of them, say Y, given the other.

Discrete case

Suppose X, Y are discrete with joint prob mass function $f_{X,Y}$.

Define the **conditional probability mass function** $f_{Y|X}(y|x)$ as

$$f_{Y|X}(y|x) := \mathbb{P}(Y = y | X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Define the **conditional expectation** of Y given $X = x$ as

$$\mathbb{E}(Y|X = x) = \sum_y y f_{Y|X}(y|x).$$

Conditional distributions (technical subtleties)

- The conditional expectation $\mathbb{E}(Y|X=x)$ is a function of x, say
$$g(x) := \mathbb{E}(Y|X=x).$$
This value is known only when we observe $X=x$.

  The expression $\mathbb{E}(Y|X)$ denotes the random variable $g(X)$.

- The definition of conditional expectation extends to continuous random variables X, Y with joint density $f_{X,Y}$.

  That is accomplished via the **conditional density function**
$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x,y)}{f_X(x)},$$
which of course is only defined for $f_X(x) > 0$.

- When $\mathbb{E}(X^2) < \infty$ the conditional expectation $g(X) := \mathbb{E}(Y|X)$ minimizes $\mathbb{E}(\|Y - g(X)\|^2)$.
In other words, $\mathbb{E}(Y|X)$ is the "projection" of Y onto the space of random variables defined by X.

Regression

Regression

Estimation of a model for the relationship between a *dependent* (or response) variable Y and an *independent* (or predictor) variable X:

$$Y = g(X) + \epsilon.$$

Regression function: $g(x) = \mathbb{E}(Y|X = x)$

Examples

- Debt payments and income in various metropolitan areas.
- Sales and advertising expenditures.
- Returns of individuals stocks and market returns.

Linear regression

Linear model assumption

$$Y = \beta_0 + \beta_1 X + \epsilon \Leftrightarrow \mathbb{E}(Y|X) = \beta_0 + \beta_1 X.$$

For a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n.$$

Assume

- $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$
- ϵ_i is uncorrelated with ϵ_j for $i \neq j$
- ϵ_i is statistically independent of X_i

Distribution parameters: $\beta_0,\ \beta_1$, and σ^2.

Interpretation of regression coefficients

β_0 : expected value of Y when $X = 0$

β_1 : expected change in Y when X increases by one unit.

Estimation of the linear regression parameters

Least squares estimation

- Population parameters: β_0 and β_1 and σ^2
- Given a set of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, we want estimators b_0 and b_1 of β_0 and β_1.
- Unobserved true error term: $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$
- Estimated error term: $e_i = Y_i - b_0 - b_1 X_i$

Choose b_0 and b_1 to minimize the estimated sum of squared errors:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

Estimation of the linear regression parameters

Least squares estimators

Some calculus and algebra shows that the solution to the least squares problem is

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \;\; b_0 = \bar{Y} - b_1\bar{X}.$$

Observe

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X \cdot S_Y} \cdot \frac{S_Y}{S_X} = r_{XY} \cdot \frac{S_Y}{S_X}$$

This shows the relationship between slope coefficient b_1 and sample correlation r_{XY}.

Least squares estimation in Excel

- Click `Data -> Data Analysis -> Regression`
- In `input Y range` highlight the data for the dependent variable
- In `input X range` highlight the data for the independent variable

We can also use `ADD TREND LINE` within a chart in Excel.

Least squares estimation in R

Suppose x and y have the relevant data. To obtain the least squares estimates use

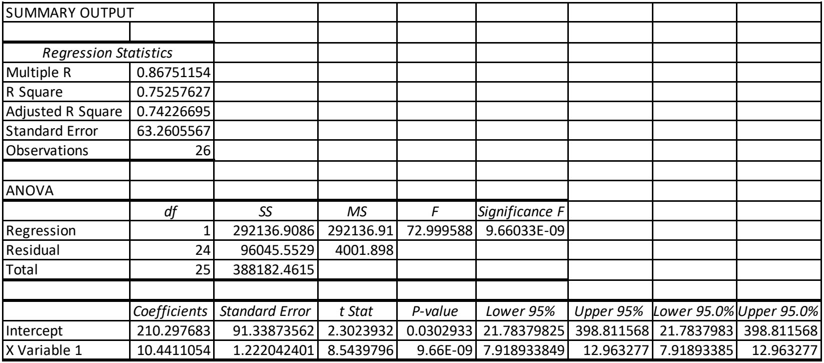
$$\texttt{lm}(\texttt{y} \sim \texttt{x})$$

Alternatively, suppose the columns x and y in the R dataframe data have the relevant data. To obtain the least squares estimates use

$$\texttt{lm}(\texttt{y} \sim \texttt{x}, \texttt{data})$$

The R command lm offers a lot more functionality but the above suffices for the purpose of simple linear regression.

Example: debt payments and median income in Excel



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.86751154							
R Square	0.75257627							
Adjusted R Square	0.74226695							
Standard Error	63.2605567							
Observations	26							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	292136.9086	292136.91	72.999588	9.66033E-09			
Residual	24	96045.5529	4001.898					
Total	25	388182.4615						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	210.297683	91.33873562	2.3023932	0.0302933	21.78379825	398.811568	21.7837983	398.811568
X Variable 1	10.4411054	1.222042401	8.5439796	9.66E-09	7.918933849	12.963277	7.91893385	12.963277

Example: debt payments and median income in R

```
Call:
lm(formula = Debt ~ Income, data = debtpayments)

Residuals:
     Min       1Q   Median       3Q      Max
-107.087  -38.767   -5.828   50.137  101.619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  210.298     91.339   2.302   0.0303 *
Income        10.441      1.222   8.544 9.66e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.26 on 24 degrees of freedom
Multiple R-squared:  0.7526,    Adjusted R-squared:  0.7423
F-statistic:    73 on 1 and 24 DF,  p-value: 9.66e-09
```

Fitted values and estimated error terms

When computing b_0, b_1 we also get:

- Fitted (predicted) values of Y:

$$\hat{Y}_i = b_0 + b_1 X_i,\ i = 1, \ldots, n$$

- Estimated error terms (residuals):

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i),\ i = 1, \ldots, n$$

- Estimator of σ^2:

$$S_e^2 = \frac{\displaystyle\sum_{i=1}^{n} e_i^2}{n-2}.$$

- Under the previous assumptions

$$\mathbb{E}(b_0) = \beta_0,\ \ \mathbb{E}(b_1) = \beta_1,\ \ \mathbb{E}(S_e^2) = \sigma^2.$$

The R^2 statistic

The sum of squares total of Y (SST) can be partitioned into:

- The portion "explained" by the regression (SSR), and
- the "unexplained" portion in the error terms (SSE).

More precisely, $SST = SSR + SSE$, where

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,\ SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,\ SSE = \sum_{i=1}^{n} e_i^2.$$

Here $\hat{Y}_i := b_0 + b_1 X_i,\ i = 1, \ldots, n$ are the "fitted values".

Coefficient of determination:

$$R^2 := \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

Interpreting the R^2 statistic

The R^2 statistic can take values between 0 and 1:

- A value of 1 would imply a perfect fit of the data to the regression line
- A value of 0 would imply that X does not explain Y

For simple regression (one predictor, case discussed so far)

$$R^2 = r_{XY}^2.$$

Confidence intervals for β_0 and β_1

Similar logic as for population mean:

- Least squares procedure gives unbiased estimators b_0 and b_1 of β_0 and β_1:

$$\mathbb{E}(b_0) = \beta_0, \;\; \mathbb{E}(b_1) = \beta_1$$

- Least squares procedure also gives estimates S_{b_0} and S_{b_1} of σ_{b_0} and σ_{b_1} respectively.

- In Excel: S_{b_0} and S_{b_1} are in the column labeled "Standard Error" adjacent to coefficient estimate.

Hypothesis testing for regression coefficients

- The logic is quite similar to that of tests for population mean.
- The most common null value is zero.
- This is a test of “statistical significance.”

To test the null hypothesis

$$H_0 : \beta_1 = 0$$

against the alternative hypothesis

$$H_1 : \beta_1 \neq 0$$

can proceed in three equivalent ways: confidence intervals, t-statistic, p-value.

Testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

First method:

Reject H_0 if

$$0 \notin b_1 \pm t_{n-2,\alpha/2} \cdot S_{b_1}$$

Second method:

Compute

$$\text{t-stat} = \frac{b_1}{S_{b_1}}$$

Reject H_0 if $|\text{t-stat}| > t_{n-2,\alpha/2}$.

Third method:

Compute p-value

$$\text{p-value} = \mathbb{P}(|t_{n-2}| > |\text{t-stat}|)$$

Reject H_0 if p-value $< \alpha$.

In least squares procedure

- The above t-stat and p-values are automatically generated.
- If we reject the null hypothesis $H_0 : \beta_1 = 0$ we say that the estimate b_1 is **statistically significant.**
- Likewise for β_0.

Intro to Prob and Stats, Week 6

Last week

- Confidence intervals
- Hypothesis testing

This week

- Regression
- The math of least-squares estimation
- Multiple linear regression
- If there is time: comparing two or more populations

Some interesting datasets we will use

- VanguardFunds: (real and recent) monthly returns of some popular mutual funds from the Vanguard family.
- StockReturns: (real and recent) excess monthly returns of some popular stocks and excess returns of the overall market.
- DebtPayments: (fictitious) data about debt payments in various metropolitan areas in the US.
- Datasets from other sources and textbooks (R, Yahoo Finance, Jaggia & Kelly's book, etc)
- Some basic data manipulation and exploration
  - Scatter plots and histograms
  - R dataframes

Regression

Regression

Estimation of a model for the relationship between a *dependent* (or response) variable Y and an *independent* (or predictor) variable X:

$$Y = g(X) + \epsilon.$$

Regression function: $g(x) = \mathbb{E}(Y|X = x)$

Examples

- Debt payments and income in various metropolitan areas.
- Sales and advertising expenditures.
- Returns of individuals stocks and market returns.

Linear regression

Linear model assumption

$$Y = \beta_0 + \beta_1 X + \epsilon \Leftrightarrow \mathbb{E}(Y|X) = \beta_0 + \beta_1 X.$$

For a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n.$$

Assume

- $\mathbb{E}(\epsilon_i) = 0$ and var$(\epsilon_i) = \sigma^2$
- ϵ_i is uncorrelated with ϵ_j for $i \neq j$
- ϵ_i is statistically independent of X_i

Distribution parameters: $\beta_0,\ \beta_1,$ and σ^2.

Interpretation of regression coefficients

β_0 : expected value of Y when $X = 0$

β_1 : expected change in Y when X increases by one unit.

Estimation of the linear regression parameters

Least squares estimation

- Population parameters: β_0 and β_1 and σ^2
- Given a set of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, we want estimators b_0 and b_1 of β_0 and β_1.
- Unobserved true error term: $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$
- Estimated error term: $e_i = Y_i - b_0 - b_1 X_i$

Choose b_0 and b_1 to minimize the estimated sum of squared errors:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

Estimation of the linear regression parameters

Least squares estimators

Some calculus and algebra shows that the solution to the least squares problem is

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \;\; b_0 = \bar{Y} - b_1\bar{X}.$$

Observe

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X \cdot S_Y} \cdot \frac{S_Y}{S_X} = r_{XY} \cdot \frac{S_Y}{S_X}$$

This shows the relationship between slope coefficient b_1 and sample correlation r_{XY}.

Least squares estimation in Excel

- Click `Data -> Data Analysis -> Regression`
- In `input Y range` highlight the data for the dependent variable
- In `input X range` highlight the data for the independent variable

We can also use `ADD TREND LINE` within a scatter plot in Excel.

Least squares estimation in R

Suppose x and y have the relevant data. To obtain the least squares estimates use

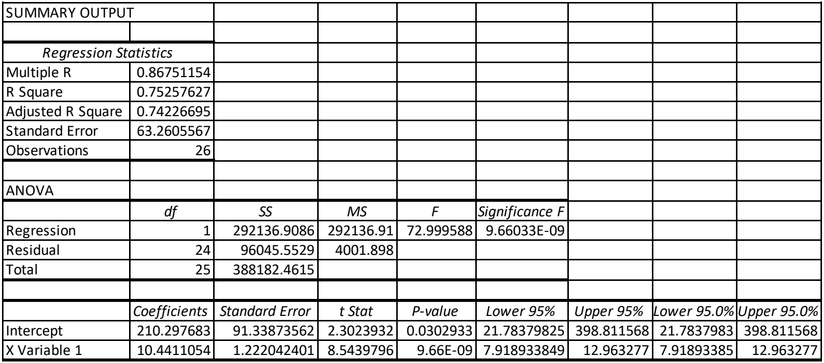
$$\mathtt{lm}(\mathtt{y} \sim \mathtt{x})$$

Alternatively, suppose the columns x and y in the R dataframe data have the relevant data. To obtain the least squares estimates use

$$\mathtt{lm}(\mathtt{y} \sim \mathtt{x}, \mathtt{data})$$

The R command lm offers a lot more functionality but the above suffices for the purpose of simple linear regression.

Example: debt payments and median income in Excel



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.86751154							
R Square	0.75257627							
Adjusted R Square	0.74226695							
Standard Error	63.2605567							
Observations	26							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	292136.9086	292136.91	72.999588	9.66033E-09			
Residual	24	96045.5529	4001.898					
Total	25	388182.4615						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	210.297683	91.33873562	2.3023932	0.0302933	21.78379825	398.811568	21.7837983	398.811568
X Variable 1	10.4411054	1.222042401	8.5439796	9.66E-09	7.918933849	12.963277	7.91893385	12.963277

Example: debt payments and median income in R

```
Call:
lm(formula = Debt ~ Income, data = debtpayments)

Residuals:
     Min       1Q   Median       3Q      Max
-107.087  -38.767   -5.828   50.137  101.619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  210.298     91.339   2.302   0.0303 *
Income        10.441      1.222   8.544 9.66e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.26 on 24 degrees of freedom
Multiple R-squared:  0.7526,    Adjusted R-squared:  0.7423
F-statistic:    73 on 1 and 24 DF,  p-value: 9.66e-09
```

Fitted values and estimated error terms

When computing b_0, b_1 we also get:

- Fitted (predicted) values of Y:

$$\hat{Y}_i = b_0 + b_1 X_i,\ i = 1, \ldots, n$$

- Estimated error terms (residuals):

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i),\ i = 1, \ldots, n$$

- Estimator of σ^2:

$$S_e^2 = \frac{\sum\limits_{i=1}^{n} e_i^2}{n-2}.$$

- Under the previous assumptions

$$\mathbb{E}(b_0) = \beta_0,\ \ \mathbb{E}(b_1) = \beta_1,\ \ \mathbb{E}(S_e^2) = \sigma^2.$$

The R^2 statistic

The sum of squares total of Y (SST) can be partitioned into:

- The portion "explained" by the regression (SSR), and
- the "unexplained" portion in the error terms (SSE).

More precisely, $SST = SSR + SSE$, where

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,\ SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,\ SSE = \sum_{i=1}^{n} e_i^2.$$

Here $\hat{Y}_i := b_0 + b_1 X_i,\ i = 1, \ldots, n$ are the "fitted values".

Coefficient of determination:

$$R^2 := \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

Interpreting the R^2 statistic

The R^2 statistic can take values between 0 and 1:

- A value of 1 would imply a perfect fit of the data to the regression line
- A value of 0 would imply that X does not explain Y

For simple regression (one predictor, case discussed so far)

$$R^2 = r_{XY}^2.$$

Confidence intervals for β_0 and β_1

Similar logic as for population mean:

- Least squares procedure gives unbiased estimators b_0 and b_1 of β_0 and β_1:

$$\mathbb{E}(b_0) = \beta_0, \ \ \mathbb{E}(b_1) = \beta_1$$

- Least squares procedure also gives estimates S_{b_0} and S_{b_1} of σ_{b_0} and σ_{b_1} respectively.
- In Excel: S_{b_0} and S_{b_1} are in the column labeled "Standard Error" adjacent to coefficient estimate.

Confidence intervals for the slope β_1 and intercept β_0

Under reasonable assumptions (ϵ_i are normal or n is large) the t-statistic

$$\frac{b_1 - \beta_1}{S_{b_1}}$$

has t-distribution with $n - 2$ degrees of freedom.

$(1 - \alpha)$ confidence interval for β_1:

$$b_1 \pm t_{n-2,\alpha/2} \cdot S_{b_1}$$

Similar for β_0.

Hypothesis testing for regression coefficients

- The logic is quite similar to that of tests for population mean.
- The most common null value is zero.
- This is a test of “statistical significance.”

To test the null hypothesis

$$H_0 : \beta_1 = 0$$

against the alternative hypothesis

$$H_1 : \beta_1 \neq 0$$

can proceed in three equivalent ways: confidence intervals, t-statistic, p-value.

Testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

First method:

Reject H_0 if

$$0 \notin b_1 \pm t_{n-2,\alpha/2} \cdot S_{b_1}$$

Second method:

Compute

$$\text{t-stat} = \frac{b_1}{S_{b_1}}$$

Reject H_0 if $|\text{t-stat}| > t_{n-2,\alpha/2}$.

Third method:

Compute p-value

$$\text{p-value} = \mathbb{P}(|t_{n-2}| > |\text{t-stat}|)$$

Reject H_0 if p-value $< \alpha$.

In least squares procedure

- The above t-stat and p-values are automatically generated.
- If we reject the null hypothesis $H_0 : \beta_1 = 0$ we say that the estimate b_1 is **statistically significant.**
- Likewise for β_0.

Example (CAPM model)

The capital asset pricing model (CAPM) states that the return R of a stock and the return of the overall market R_M are related via

$$R - R_f = \alpha + \beta \cdot (R_M - R_f) + \epsilon,$$

where R_f is the risk-free return.

The CAPM also postulates that $\alpha = 0$.

In other words, the CAPM states the null hypothesis

$$H_0 : \alpha = 0.$$

Use data to test the validity of the CAPM on various stocks.

Other hypothesis tests

To test for a different null value $H_0 : \beta_1 = \bar{\beta}$, compute instead

$$\text{t-stat} = \frac{b_1 - \bar{\beta}}{S_{b_1}}$$

Example

In the debt payments example test the null hypothesis

$$H_0 : \beta_1 = 11$$

against the alternative $H_1 : \beta_1 \neq 11$ at the 5% significance level

To test a one-sided hypothesis:

$$H_0 : \beta_1 \leq \bar{\beta} \quad \text{or} \quad H_0 : \beta_1 \geq \bar{\beta},$$

proceed as before but reject only on one of the tails.

Example 1

In CAPM example for Apple, test null hypothesis

$$H_0 : \beta \leq 1$$

against the alternative $H_1 : \beta > 1$ at the 5% significance level.

Example 2

In CAPM example for Johnson & Johnson, test null hypothesis

$$H_0 : \beta \geq 1$$

against the alternative $H_1 : \beta < 1$ at the 5% significance level.

Prediction and Prediction Intervals

Back to the debt payments example

Predict debt payments if income is \$80,000.

This is a new observation $(n+1)$ where we do not know Y_{n+1} yet.

Predicted value

$$\hat{Y}_{n+1} = b_0 + b_1 X_{n+1}$$

Prediction interval

The above predicted value is only a point estimate. To obtain a $(1-\alpha)$ prediction interval compute: $\hat{Y}_{n+1} \pm t_{n-2,\alpha/2} \cdot S_p$.

Here S_p is the "standard *prediction* error":

$$S_p = S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

Transformed variables

Sometimes the relationship between Y and X is nonlinear but it can be recast as a linear one using a suitable transformation.

Examples

- Log-log models: Regress $\log(Y)$ on $\log(X)$.
- Semi-log models: Regress $\log(Y)$ on X.

Example (Moore's law)

- In 1965, Gordon Moore, an Intel vice president made the prediction that transistors per chip would double every year. In 1975 we revised his forecast to doubling every two years.
- This has come to be known as Moore's Law.
- How can we formulate a regression model to test this prediction?
- We show that a semi-log model is the appropriate form.

The semi-log form is:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon$$

To test Moore's law, I obtained data on the transistor count for Intel chips produced since 1971.
(See Transistor count Wikipedia page.)

For each new chip, I recorded:

- Transistors per chip (dependent variable Y)
- Date each chip was introduced
- Independent variable X: time in months with month 1 being January 1971.

Moore's law corresponds to the null hypothesis

$$H_0 : 24\beta_1 = \log(2)$$

or equivalently

$$H_0 : \beta_1 = \frac{\log(2)}{24}.$$

What does the data say?

The math of least-squares estimation

Linear model assumptions again

We have a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n$$

and

- $\mathbb{E}(\epsilon_i) = 0$ and var$(\epsilon_i) = \sigma^2$
- ϵ_i is uncorrelated with ϵ_j for $i \neq j$
- ϵ_i is independent of X_i

The math of least-squares estimation

Let $\mathbf{b} \in \mathbb{R}^2, \mathbf{Y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times n}$ be as follows

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

The least-squares estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ is

$$\min_{\mathbf{b}} (\mathbf{Y} - \mathbf{X}\mathbf{b})^{\mathsf{T}} (\mathbf{Y} - \mathbf{X}\mathbf{b}) \rightsquigarrow \mathbf{b} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}.$$

Theorem (sampling distribution of $\mathbf{b}$)

Under the above linear model assumptions the least-squares estimator $\mathbf{b}$ *is unbiased*

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta},$$

and has the following variance-covariance matrix

$$\sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}.$$

Standard errors of b_0, b_1

The diagonal entries of $\sigma^2 \cdot (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ give the standard errors of b_0, b_1:

$$\mathsf{se}(b_0) = \sigma \cdot \sqrt{\frac{\sum_{i=1}^n X_i^2}{n \cdot \sum_{i=1}^n (X_i - \bar{X})^2}}$$

and

$$\mathsf{se}(b_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

The least-squares procedure generates estimates of these standard errors via the following unbiased estimator of σ^2:

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2.$$

More precisely, the least-squares procedures in R, Excel, etc compute the following estimates of $\mathsf{se}(b_0)$ and $\mathsf{se}(b_1)$ respectively

$$S_{b_0} = S_e \cdot \sqrt{\frac{\sum_{i=1}^n X_i^2}{n \cdot \sum_{i=1}^n (X_i - \bar{X})^2}} \quad \text{and} \quad S_{b_1} = \frac{S_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Inference

Suppose the residuals ϵ_i are normal. Then for $i = 0, 1$:

$$\frac{b_i - \beta_i}{S_{b_i}} \sim t_{n-2}.$$

This is why we can perform hypothesis testing on β_0 and β_1.

When the linear model assumptions hold, the least-squares estimator is the "best linear unbiased estimator" (BLUE).

Theorem (Gauss-Markov Theorem)

Suppose the linear model assumptions hold. Then the least-squares estimator has the lowest variance among the class of unbiased linear estimators of $\boldsymbol{\beta}$.

What if the linear model assumptions do not hold?

How to check that linear model assumptions hold?

- Plot of residuals should show no evident pattern.
- Formal tests.

What to do if the linear model assumptions do not hold?

- Variable transformations.
- Modifications to the least-squares procedure.
- Other types of regression analysis.

Residual plots

When the linear model assumptions hold, the plot of e_i (residuals) versus X_i (explanatory variable) should show no pattern.

Violation 1: non-linearity

When the residual plot suggests a nonlinear relationship, e.g., quadratic.

In this case a variable transformation could transform the problem into one where the linear assumptions hold.

Violation 2: heteroskedasticity

When the residual plot suggests that the variance of the residuals depends on the explanatory variables.

In this case the estimation of the standard errors is biased. Consequently the inference via confidence intervals, t-statistics, and p-values is not reliable.

There are two possible ways of dealing with heteroskedasticity:

- Use “weighted least squares” to incorporate the dependence between the residual variance and the explanatory variable.
- Use a “robust standard error estimation”. This is often referred to as the “White correction”.

Violation 3: residuals are correlated

This often happens with time-series data. Again in this case the estimation of the standard errors is biased. A popular remedy is the “Newey-West” correction to the standard errors estimation.

An alternative is to use more elaborate techniques for time-series data.

Multiple linear regression

Multiple linear regression

Linear model assumption

$$Y = \beta_0 + \beta_1 X_1 + \cdots + X_p + \epsilon.$$

For a random sample $(X_{11}, \ldots, X_{1p}, Y_1), \ldots, (X_{n1}, \ldots, X_{np}, Y_n)$ we have:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i,\ i = 1, \ldots, n.$$

Assume

- $\mathbb{E}(\epsilon_i) = 0$ and var$(\epsilon_i) = \sigma^2$
- ϵ_i is uncorrelated with ϵ_j for $i \neq j$
- ϵ_i is statistically independent of $X_{i1}, \ldots, X_{ip}$

The math of least-squares estimation again

Let $\mathbf{b} \in \mathbb{R}^{p+1}, \mathbf{Y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ be as follows

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & & \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}.$$

The least-squares estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ is

$$\min_{\mathbf{b}} (\mathbf{Y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{Y} - \mathbf{X}\mathbf{b}) \rightsquigarrow \mathbf{b} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}.$$

Theorem (sampling distribution of $\mathbf{b}$)

Under the above linear model assumptions the least-squares estimator $\mathbf{b}$ *is unbiased*

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta},$$

and has the following variance-covariance matrix

$$\sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}.$$

We also have the following unbiased estimator of σ^2:

$$S_e^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\mathbf{Y}-\hat{\mathbf{Y}})_i^2.$$

where $\hat{\mathbf{Y}} := \mathbf{X}^\mathsf{T}\mathbf{b}$ are the fitted values.

The diagonal entries of $\mathbf{S}_e^2 \cdot (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ give unbiased estimators $S_{b_i}^2$ of the variances of b_i for $i = 0, 1, \ldots, p$.

Again we have $SST = SSR + SSE$ where

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,\ SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,\ SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Least squares estimation in Excel and R

In Excel: same as before. “X range” should include all predictors.
In R:

$$\texttt{lm}(\texttt{y} \sim \texttt{x1} + \cdots + \texttt{xp}, \texttt{data})$$

Inference

Suppose the residuals ϵ_i are normal. Then for $i = 0, 1, \ldots, p$:

$$\frac{b_i - \beta_i}{S_{b_i}} \sim t_{n-p-1}.$$

Get confidence intervals, t-stats, p-values.

When the linear model assumptions hold, the least-squares estimator is the "best linear unbiased estimator" (BLUE).

Theorem (Gauss-Markov Theorem)

Suppose the linear model assumptions hold. Then the least-squares estimator has the lowest variance among the class of unbiased linear estimators of $\boldsymbol{\beta}$.

Joint hypothesis testing in multiple linear regression

To test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
against $H_1 : \beta_i \neq 0$ for some $i = 1, \ldots, p$ proceed as follows.

Compute the F-statistic

$$\text{F-stat} := \frac{SSR/p}{SSE/(n-p-1)}.$$

Under the null hypothesis F has F-distribution with parameters p and $n-p-1$.

Rule: reject H_0 if F-stat $> F_{p,n-p-1,\alpha}$.

F distribution

Suppose S_1 and S_2 are independent chi-square with df1 and df2.
Then $\dfrac{S_1/\text{df}1}{S_2/\text{df}2}$ has F-distribution with parameters $(\text{df}1, \text{df}2)$.

Challenges in multiple linear regression

Multicollinearity

When one of the predictors is nearly a linear transformation of the others. In this situation the estimates of the coefficients is inaccurate. It is difficult to disentangle the separate effects of the predictor variables.

Model selection

Suppose there are many predictors, that is, p is large. Which predictors should we include?
Too few predictors yield high bias. This is "underfitting".
Too many predictors yield high variance. This is "overfitting".
A popular model to deal with this challenge: Lasso.

If there is time: comparing two or more populations

Comparing two populations

Confidence intervals for difference of population means

(measured difference) ± (MOE)

Two common situations

- Matched (paired) samples
- Independent samples

We can also test null hypotheses about differences in the population means.

Confidence interval using matched samples

Suppose we have a random sample of n matched pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ for the variables X and Y.

Put $d_i := X_i - Y_i,\ i = 1, \ldots, n$.

The $(1 - \alpha)$ confidence interval for $\mu_X - \mu_Y$ is:

$$\bar{d} \pm t_{n-1,\alpha/2} \cdot \frac{S_d}{\sqrt{n}},$$

where S_d^2 is the sample variance of $d = X - Y$.

Confidence interval using independent samples

Suppose we have two independent random samples of $X_1, X_2, \ldots, X_{n_X}$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ for the variables X and Y.

Confidence interval for $\mu_X - \mu_Y$

$$(\bar{X} - \bar{Y}) + \text{MOE}.$$

The MOE depends on our assumptions on the population variances.

Independent samples with known population variances

The $(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is:

$$(\bar{X}-\bar{Y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

Independent samples & unknown equal variances

$(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$:

$$(\bar{X}-\bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2} \cdot \sqrt{\frac{S_p^2}{n_X} + \frac{S_p^2}{n_Y}}$$

for the "pooled" variance"

$$S_p^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}$$

Difference of two population proportions

Let $\hat{p}_X$ and $\hat{p}_Y$ be the sample proportions for samples of size n_X and n_Y drawn from populations with proportions p_X and p_Y.

The $(1-\alpha)$-confidence interval for the $p_X - p_Y$ is:

$$\hat{p}_X - \hat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$$

Example

On May 23, 2003, the New York Times reported an important side effect of a clinical trial of the drug Prempro.

In the trial, 40 of the subjects receiving the medication developed dementia while 21 of the subjects receiving a placebo developed dementia.

There were a total of 4,532 subjects, with an equal number in the group receiving medication and the group receiving the placebo.

- Compute a 95% confidence interval for the difference of population proportions.
- Would you conclude that the difference in frequencies of dementia between the two groups was probably due to chance?

Comparing variances

Suppose we have two independent random samples of $X_1, X_2, \ldots, X_{n_X}$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ from normal distributions with the same variance.
Then the F-statistic

$$\text{F-stat} = \frac{S_X^2}{S_Y^2}$$

has F-distribution with parameters $(n_X - 1, n_Y - 1)$.

F distribution

Suppose S_1 and S_2 are independent chi-square with df1 and df2.
Then $\frac{S_1/\text{df}1}{S_2/\text{df}2}$ has F-distribution with parameters $(\text{df}1, \text{df}2)$.

One-sided test of variance

To test $H_0 : \sigma_X^2 \leq \sigma_Y^2$.
Rule: reject H_0 if F-stat $> F_{n_X-1, n_Y-1, \alpha}$.

Analysis of Variance (ANOVA)

One-way ANOVA is a simple technique to compare the mean of several populations.

It relies on the assumption that the populations have the same variance. This is a similar assumption to that in linear regression.

Indeed, one-way ANOVA can be seen as a special case of multiple regression via "dummy" variables.

One-way ANOVA setup

- There are p populations we are interested in comparing.
- For population $i = 1, \ldots, p$ we have a random sample $Y_{i1}, \ldots, Y_{in_i}$.

Sample means

Sample mean for population i

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}.$$

Overall sample mean

$$\bar{Y} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}}{n} = \frac{\sum_{i=1}^{p} n_i \bar{Y}_i}{n}.$$

Sum of squares decomposition

Define the following sums of squares.

Within-groups

$$SSW = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Between-groups

$$SSG = \sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2$$

Total

$$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$$

Sum of squares decomposition: $SST = SSW + SSG$.

ANOVA for comparing multiple populations

To test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

against the alternative

$$H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j.$$

Compute the F-statistic

$$\text{F-stat} := \frac{SSG/(p-1)}{SSW/(n-p)}$$

Under the null hypothesis, F-stat has F-distribution with parameters $p-1$ and $n-p$.

Rule: reject H_0 if F-stat $> F_{p-1,n-p,\alpha}$.

ANOVA in linear regression

A similar analysis of variance applies to linear regression.
To test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ proceed as follows.
Compute the F-statistic

$$\text{F-stat} := \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{S_e^2}.$$

Again under the null hypothesis F has F-distribution with parameters 1 and $n-2$.

Rule: reject H_0 if F-stat $> F_{1,n-2,\alpha}$.

Neat connection with t-test

We have F-stat $=$ t-stat^2 and the above rule is identical to the rule: reject if $|\text{t-stat}| > t_{n-2,\alpha/2}$.