MATH 60604A Statistical Modelling

Ch1: Introduction & Review

HEC Montréal
Department of Decision Sciences

Chapter Overview

1. Review: Probability
2. Review: Random Variables
   - Discrete Distributions
   - Continuous Distributions
   - Moments
3. Review: Inference
   - Basic Notions
   - Sampling and Estimation
   - Hypothesis Tests
   - Confidence Intervals
4. Concluding remarks

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Probability basics

- Let A and B denote arbitrary events, $\mathcal{S}$ the sample space, and let $P(\cdot)$ denote the probability of, then
  - $P(\mathcal{S}) = 1$
  - $P(A) \in [0, 1]$
  - $P(A^{\complement}) = 1 - P(A)$

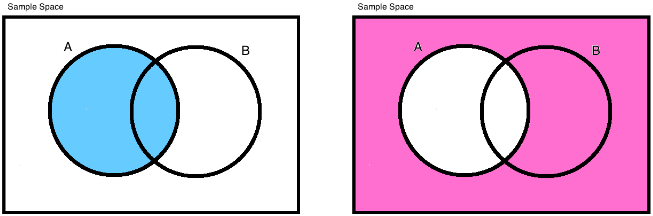


Figure: left: event A, right: event $A^{\complement}$ (complement of A)

Probability basics

- Unions and intersections:

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

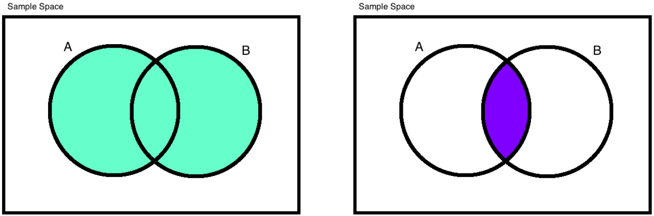


Figure: left: event $A \cup B$ (A union B), right: event $A \cap B$ (A intersection B)

Probability basics

- Law of total probability

$$P(A) = P(A \cap B) + P(A \cap B^{\complement})$$

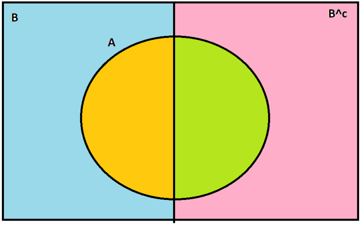


Figure: yellow: $A \cap B$, green: $A \cap B^{\complement}$

- in more general terms, for mutually exclusive events $E_1, E_2, \ldots, E_K$ that partition the sample space, i.e. such that $P\left(\cup_{i=1}^{K} E_i\right) = \sum_{i=1}^{K} P(E_i) = 1$, we have that

$$P(A) = \sum_{i=1}^{K} P(A \cap E_i)$$

Review: Probability
Review: Random Variables
Discrete Distributions
Continuous Distributions
Moments
Review: Inference
Basic Notions
Sampling and Estimation
Hypothesis Tests
Confidence Intervals
Concluding remarks

Probability basics

- Conditional probability: the probability of A given B

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

- Bayes' Theorem: reformulating conditional probability

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^{\complement})P(A^{\complement})}$$

  - simply combining the definition of conditional probability and the law of total probability!

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Random Variables

- A random variable is a mapping from the sample space to a real number $X : \mathcal{S} \mapsto \mathbb{R}$
  - ex: flipping a coin, define $X = 0$ if it lands on tails, $X = 1$ if it lands on heads
  - ex: rolling two dice, random variable of interest X is the sum of the two outcomes

outcome	value x of X	$P(X = x)$
$(1, 1)$	2	$1/36$
$(1, 2), (2, 1)$	3	$2/36$
$(1, 3), (2, 2), (3, 1)$	4	$3/36$
$(1, 4), (2, 3), (3, 2), (4, 1)$	5	$4/36$
$(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$	6	$5/36$
$(1, 6), (2, 5), (3, 4)(4, 3), (5, 2), (6, 1)$	7	$6/36$
$(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$	8	$5/36$
$(3, 6), (4, 5), (5, 4), (6, 3)$	9	$4/36$
$(4, 6), (5, 5), (6, 4)$	10	$3/36$
$(6, 5), (5, 6)$	11	$2/36$
$(6, 6)$	12	$1/36$

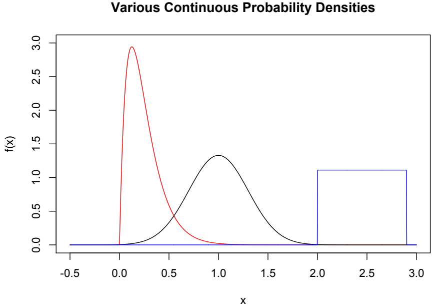
  - Note: an uppercase X is used to denote the random variable while lowercase x is used to denote a *realisation* of the random variable (i.e. an actual value it takes on).

Random Variables

- Types of random variables:
  - discrete random variables: can only take on a countable number of possible values (possibly infinite, but countably so!)
    - ex: $X =$ the number of times a coin lands on heads in 10 tosses, $X \in \{0, 1, 2, \ldots, 10\}$
    - ex: $X =$ the number of tosses until a coin lands on heads, $X \in \{1, 2, 3, \ldots\}$
    - ex: $X =$ the number of penalties in a hockey game $X \in \{0, 1, 2, 3, \ldots\}$
  - continuous random variables: can take on any real number in some interval (there are thus uncountably many possible values, i.e. to any amount of decimal places)
    - ex: $X =$ the time (in minutes) it takes a student to finish their final exam, $X \in [0, 180]$
    - ex: $X =$ the total amount of winnings at a casino (in \$), $X \in (-\infty, \infty)$
    - ex: $X =$ the time (in minutes) until the next bus arrives, $X \in [0, \infty)$

Random Variables

- The probability distribution of a random variable X is a function or rule that assigns probabilities to the different values that X can take on.
  - discrete random variables: the probability distribution, also called probability mass function, is denoted by $P(X = x)$
    - ex: tossing a (fair) coin, let $X = 0$ if the coin lands on tails and $X = 1$ is it lands on heads, then $P(X = 0) = P(X = 1) = 1/2$
  - continuous random variables: the probability distribution, also called the probability density function or simply density, is a smooth curve, denoted by $f(x)$, some examples are shown in the Figure below



Random Variables

Some remarks on continuous random variables...

- Recall: a continuous random variable X can take on *any* real value in an interval, so rather than defining the probability that $X = x$ point by point (impossible!) we consider the probability that X falls in a given interval, e.g.

  $$P(a \leq X \leq b), \quad P(X \leq a), \quad P(X \geq b)$$

  (where a and b are some arbitrary values, e.g. $a = 10$ and $b = 200$)

- the density $f(x)$ of a continuous random variable X allows to assign probabilities over intervals as the area under the curve $f(x)$:

  $$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  - Note: because a continuous random variable X can take *any* value in a given interval, $P(X = x) = 0$ for any given point x. This implies that:

    $$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

    **careful, this is not true for discrete random variables

Some basic distributions

- There are many well known probability distributions that are used very often to model data.
- These distributions are defined in terms of **parameters**
  - family of distributions: general "recipe" for the distribution, in terms of arbitrary parameter values
    - ex: the Normal distribution $\mathcal{N}(\mu, \sigma^2)$
  - specific distribution: actual "cake" for the distribution, in terms of specific parameter values (so that can compute things!)
    - ex: the standard Normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 1)$
  - modelling data: finding “good" estimates for the parameter values (given an appropriate distribution) so that we can make probabilistic statements involving the phenomena under investigation
    - ex: modelling how much money people make gambling at a casino, assume a normal distribution $\mathcal{N}$ with $\hat{\mu} = -20$ and $\hat{\sigma} = 10$
- We will revisit some of these distributions within a regression framework

Table of Contents

1. Review: Probability
2. Review: Random Variables
   - Discrete Distributions
   - Continuous Distributions
   - Moments
3. Review: Inference
   - Basic Notions
   - Sampling and Estimation
   - Hypothesis Tests
   - Confidence Intervals
4. Concluding remarks

Discrete Distributions

Bernoulli distribution:

- A Bernoulli random variable X takes on the values 0 or 1
  - useful for modelling dichotomous outcomes:
    - $X = 1 =$ “success”, $X = 0 =$ “failure”
    - $X = 1 =$ “yes”, $X = 0 =$ “no”
    - $X = 1 =$ “present”, $X = 0 =$ “absent”
- The Bernoulli distribution is parametrized in terms of p *(think of p as the “probability of success”, whatever we consider success to be)*:

$$X \sim \ Bernoulli(p), \quad X \in \{0, 1\}$$

and the probability distribution is given by

$$P(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Discrete Distributions

Binomial distribution:

- The Binomial distribution is parametrized in terms of n and p
- General binomial set up:
  - n independent trials;
  - each trial results in either a "success" or "failure"
  - the probability of a "success" is p for each trial (identical trials)
  - Define the random variable Y to be the **number of successes in the n trials**, $Y \in \{0, 1, \ldots, n\}$
- $Y \sim$ *Binomial*(n, p) and has probability distribution given by

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

where

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

Discrete Distributions

Binomial distribution:

- Note: $Y \sim \ \mathit{Binomial}(n, p)$ if

$$Y = \sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

  where the X_i's are independent and identically distributed Bernoulli random variables with $X_i \sim \ \mathit{Bernoulli}(p)$

- In other words, a Binomial random variable is just the sum of n independent Bernoulli random variables each with the same probability of "success" p

Discrete Distributions

Poisson distribution

- A Poisson random variable X takes on the values $\{0, 1, 2, \ldots\}$
  - useful for modelling count data
    - ex: $X =$ the number of BIXI rentals on a given day
    - ex: $X =$ the number of insurance claims a policyholder makes in a year
    - ex: $X =$ the number of COVID cases in Montreal
- The Poisson distribution is parametrized in terms of λ

$$X \sim \ Poisson(\lambda), \quad X \in \{0, 1, 2, \ldots\}$$

and the probability distribution is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Discrete Distributions

Negative binomial distribution

- A negative binomial random variable X takes on the values $\{0, 1, 2, \dots\}$ (sometimes on $\{1, 2, 3, \dots\}$, depending on the parametrization considered)
  - useful for modelling count data
- The negative binomial distribution is parametrized in terms of $r \in \mathbb{N}$ and $p \in (0, 1)$ (other parametrizations possible)
  - X: number of failures until r^{th} success, where the probability of a success is p; the distribution function is

$$P(X = x) = \binom{x + r - 1}{x} (1 - p)^x p^r, \qquad x \in \{0, 1, 2, \dots\}$$

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

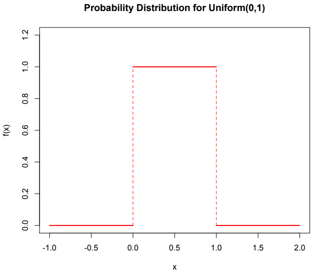
3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Continuous Distributions

Uniform distribution

- A uniform r.v. X can take on any value in a given interval (a, b)
- It is the simplest "type" of continuous random variable:
  - its density $f(x)$ is flat, that is, it assigns equal probabilities to intervals of the same width
  - the parameters a and b represent the interval on which the density is defined (i.e. the minimum and maximum possible values)
  - ex: $\mathcal{U}(0, 1)$:



Continuous Distributions

Uniform distribution

- The uniform distribution is parametrized in terms of a and b,

$$X \sim \mathcal{U}(a, b)$$

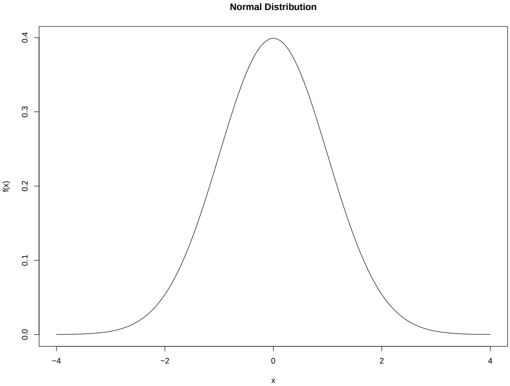
  and has density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for all } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Continuous Distributions

Normal distribution

- A normal (or Gaussian) random variable X can take any real value
    - it is probably the most well known distribution (bell shape curve)!



    - the normal distribution plays a key role in statistical inference, as we will see later

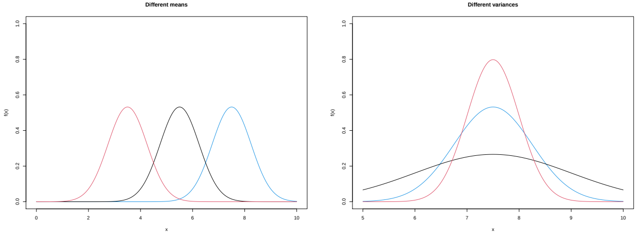
Continuous Distributions

Normal distribution

- The normal distribution is parametrized in terms of μ and σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, and has density

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

  - μ: the mean → center of the density (location)
  - σ^2: the variance → shape of the density (scale)



Continuous Distributions

Normal distribution

- We often work with the "standard normal" distribution, which is the special case where $\mu = 0$ and $\sigma^2 = 1$, usually denoted by $Z \sim \mathcal{N}(0,1)$
- The normal distribution has many "nice properties", ex:
  - its density is symmetric about its mean
  - the mean = median = mode
  - if $X = \mu + \sigma Z$, then $X \sim \mathcal{N}(\mu, \sigma^2)$, or, equivalently,

    if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$

    (we often refer to $Z = (X - \mu)/\sigma$ as a *z score*)
  - much more...
- we often use the normal distribution to make approximations for other distributions (that may too complex to work with)
  - can use the normal distribution to approximate the binomial distribution when n is "large"

Continuous Distributions

Normal distribution

- It's difficult to computer probabilities for the normal distribution (in general, the integral is not available in closed form), so rather we use standard normal tables or a statistical software

- some useful probabilities (come up again with confidence intervals...)
  - $P(X \in \mu \pm \sigma) = P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$
  - $P(X \in \mu \pm 2\sigma) = P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$
  - $P(X \in \mu \pm 3\sigma) = P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$

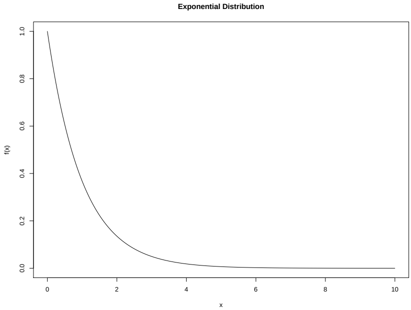
Continuous Distributions

Exponential distribution

- An exponential r.v. X can take on any positive value, in $(0, \infty)$
- Its density involves a single parameter β

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

- ex: $Exp(1)$



- The parameter β is also the mean

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Moments

- We are often interested in the moments of a random variable X:
  - mean: the mean (or expected value or first moment) is

$$\begin{aligned} \mu = \mathrm{E}(X) &= \sum_x x\, P(X = x) && \text{for a discrete r.v.} \\ &= \int_x x\, f(x)\, dx && \text{for a continuous r.v.} \end{aligned}$$

think of it as weighting all possible outcomes by their respective probability

    - the mean is a measure of central tendency

  - variance: the variance (second central moment, about the mean) is

$$\begin{aligned} \sigma^2 = \mathrm{var}(X) &= \sum_x (x-\mu)^2 P(X = x) && \text{for a discrete r.v.} \\ &= \int_x (x-\mu)^2 f(x)\, dx && \text{for a continuous r.v.} \end{aligned}$$

    - the variance is a measure of dispersion

Table of Contents

1. Review: Probability
2. Review: Random Variables
   - Discrete Distributions
   - Continuous Distributions
   - Moments
3. Review: Inference
   - Basic Notions
   - Sampling and Estimation
   - Hypothesis Tests
   - Confidence Intervals
4. Concluding remarks

Table of Contents

1 Review: Probability

2 Review: Random Variables

Discrete Distributions

Continuous Distributions

Moments

3 Review: Inference

Basic Notions

Sampling and Estimation

Hypothesis Tests

Confidence Intervals

4 Concluding remarks

Population

- **Population** of interest: the collection of units (e.g. individuals, objects, events, etc.) that we are interested in studying:
  - ex: residents of Quebec, graduate students at HEC, etc.
- generally, it is too difficult or costly to observe the entire population
  - exception: census (when collect data/information on all units in a population)
- rather, we observe a sample from the population
  - gives us some information about the overall population
- **Statistical inference** seeks to make conclusions on the whole population (*infer*) using the information in the sample
  - the idea is that we observe a sample of units on which we can measure some characteristic(s) of interest, and we use the information in the sample to infer the characteristics of the whole population of interest
  - we want to generalize a conclusion from a sample to a larger population

Sample

- **Sample:** subset of units (e.g. individuals) from the population that are actually observed
  - it's the actual data we collect as part of our "experiment"
  - we analyze the sample in order to make relevant conclusions pertaining to the population of interest
- In order to make relevant conclusions, it is essential to have a "good" sample; the sample must be
  - representative of the population being studied (i.e. the composition of the sample must be similar to that of the population)
  - sufficiently large
  - accurately measured (i.e. we must properly measure the characteristic(s) of interest on the sample units)
- There are entire courses dedicated to sampling theory!

Variables

- **Variables** are the individual characteristics of the population that we are interested in studying
  - ex: age, salary, consumer habits, etc.
  - a variable or characteristic varies from one unit (or individual) to another
  - we often denote variables by X or Y
- Different types of variables:
  - quantitative: represent a quantity, numerical in nature
    - ex: salary, temperature, number of visits to a website, GPA, etc.
  - qualitative: represent a quality, categorical in nature
    - ex: country of origin, letter grade, pass/fail, etc.
  - it is very important to be able to distinguish between types of variables as this will influence the choice of model and analysis that is carried out

Variables

- Quantitative variables
  - continuous: can take on any value in an interval i.e. there are an infinite number of possible values
    - ex: time, height, weight, speed, salary
  - discrete: can take on countably (possible infinite) many possible values
    - number of visits to a website, number of items bought at a store
    - can sometimes be approximated by a continuous variable...
- Qualitative variables
  - nominal variables: no ordering among categories
    - ex: country of origin, type of product bought in a grocery store, etc.
  - ordinal variables: categories are ordered
    - annual salary categorized as (<20000, 20000 to 40000, 40000 to 60000, >60000)
    - satisfaction with a service (not satisfied, somewhat satisfied, satisfied, very satisfied, extremely satisfied)

Qualitative variables

- Generally, qualitative variables must be coded numerically so that statistical analyses can be carried out
- nominal variables
  - numerical coding is completely arbitrary
  - ex: country of origin 1=Canada, 2=USA, 3=Mexico, etc.
  - ex: type of product 1=fruits, 2=vegetables, 3=fish, etc.
- ordinal variables
  - numerical coding is meaningful and must respect the order inherent in the categories; it is possible to treat these variables as quantitative under certain conditions
  - ex: satisfaction: 1=not satisfied, 2=somewhat satisfied, 3=satisfied, 4=very satisfied, 5=extremely satisfied
  - ex: annual salary: 1=<20000, 2=20000 to 40000, 3=40000 to 60000, 4=>60000

More terminology

- In general, a statistical model is used to explain some phenomenon that we are interested in studying
  - ex: we could use a Poisson distribution to model the number of daily BIXI rentals at a given station
- Typically, the model is defined in terms of parameters
  - recall: the parameters of a distribution allow us to calculate probabilities and other interesting quantities
  - ex: the Poisson distribution is parametrized in terms of λ; if we know the value of λ we have all the information we need to compute any probability, moment, etc! In the BIXI problem, we could calculate the mean number of daily bike rentals, or the probability that there are more than 20 rentals taken from a particular station, etc.

More terminology

- **Parameter:** quantity associated with (the distribution of) a random variable X at the population level ("truth")
  - the parameter is considered fixed, but unknown
    - if the parameter value was known, we would have all the information necessary to describe the phenomenon under question, and there would be nothing to investigate!
    - note: this is a *frequentist* statistical approach, in a "Bayesian" framework, we treat the parameters as random variables themselves!
  - often, the parameter value is what motivates the study to begin with
    - the idea is that knowledge of the parameter value allows us to describe (via a statistical model) the phenomenon under question
    - ex: the average daily BIXI rentals λ at a particular station
    - ex: the average error μ made when refunding invoices

More terminology

- **Estimation:** is the evaluation (estimation) of an unknown parameter based on an observed sample
  - "best guess" or approximation of the true value based on observed data
  - the estimate depends on the sample used to calculate it
    - $\Rightarrow$ an estimate will vary according to the actual observed sample
    - this is referred to as **sampling variability**
  - *note that the actual sample in itself is not of interest, it is merely a study tool used to understand the underlying population of interest and ultimately estimate the parameter of interest*

Table of Contents

1. Review: Probability
2. Review: Random Variables
    - Discrete Distributions
    - Continuous Distributions
    - Moments
3. Review: Inference
    - Basic Notions
    - Sampling and Estimation
    - Hypothesis Tests
    - Confidence Intervals
4. Concluding remarks

More terminology

- Random sample: sample of data points from the population of interest where each observation is *independent* and *identically distributed*
  - many of the statistical methods we will see are based on the assumption that the data used form a random sample, that is, observations are i.i.d. (independent and identically distributed)
  - ex: suppose we're interested in investigating the amount of time students from HEC spend studying for their final exam, which of the following would consist of a "good" sample?
    - sample 1: take the students in this class as the sample
    - sample 2: randomly select students in the MSc Data Science and Business Analytics program
    - sample 3: randomly select students across all programs

The mean

- Often the parameter of interest is the population mean, μ
  - an intuitive way to estimate the population mean based on observed sample data is to simply calculate the average
- Sample mean: for a (random) sample $X_1, X_2, \ldots, X_n$, the sample mean is

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

the sample mean is simply the average of the sample data

- Recall: sampling variability
  - $\bar{X}$ depends on the sample, that is, it will vary from sample to sample
  - $\bar{X}$ is itself a random variable (whose distribution will depend on the distribution of the sample $X_1, \ldots, X_n$)

Sample mean

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

- $\bar{X}$ is a random variable
  - X_i : pre-experiment, before we actually collect data and measure our variables, everything in terms of random variables $X_1, X_2, \ldots, X_n$
  - x_i : post-experiment, after we actually collect data and measure our variables, we have actual numbers to work with $x_1, x_2, \ldots, x_n$
  - Ex: we want to study the daily average amount of hand sanitizer (in L) used at CLSC Côte-des-neiges. To do so, we collect data at random for $n = 100$ days.
    - pre-experiment: $X_1, \ldots, X_{100}$ represent the *random* quantities and $\bar{X}$ represents the *random* sample mean
    - post-experiment: we have actual data, e.g. $x_1 = 200, x_2 = 250, \ldots, x_{100} = 187$ and we can calculate $\bar{x} = 199.6$

Sample mean

- For a random sample $X_1, \dots, X_n$ (that is, i.i.d.) where each X_i has mean μ and variance σ^2, the sample mean has:
  - mean μ:

    $$\boxed{\mathrm{E}(\bar{X}) = \mu}$$

    since

    $$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

  - variance σ^2/n:

    $$\boxed{\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}}$$

    since

    $$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

    notice: what happens to $\mathrm{var}(\bar{X})$ as $n \to \infty$, that is, as the sample size gets very large?

The Central Limit Theorem

Central Limit Theorem (CLT)

If $X_1, \ldots, X_d$ are i.i.d. with $\mathrm{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2 < \infty$ then for large n,

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- The CLT is a very important theorem that plays an essential role in inferential statistics
- Remarks:
  - notice that we don't need to know the exact distribution of the individual X_i, only that they are i.i.d. with finite variance!
  - the CLT is an asymptotic statement, for $n \to \infty$; when n is "small", we can still sometimes find the distribution of $\bar{X}$ by hand, but it is not always straightforward, ex: $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
  - in practice, we often don't know the true value of σ^2, so we estimate it be the sample variance $\hat{\sigma}^2 = S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$

Table of Contents

1 Review: Probability

2 Review: Random Variables

Discrete Distributions

Continuous Distributions

Moments

3 Review: Inference

Basic Notions

Sampling and Estimation

Hypothesis Tests

Confidence Intervals

4 Concluding remarks

Hypothesis testing

- A statistical hypothesis test is a way to evaluate the statistical evidence provided by a sample in order to make a decision regarding the underlying population (inference)
- There are four main “steps” involved in a hypothesis test:
  - (1) define the hypotheses being tested
    - (1a) carry out experiment / collect sample data
  - (2) calculate the test statistic
  - (3) calculate the p-value (or rejection region)
  - (4) make a conclusion

Hypothesis testing

- The general framework for hypothesis testing is always similar, following those 4 main steps. The exact details will change depending on the context / research question
  - ex: one sample t-test for the mean
  - ex: two independent samples t-test for the mean
  - ex: paired t-test for the mean
  - ex: chi-squared test for independence
  - ex: F-test for two population variances
  - etc!

1) Hypotheses

- Statistical tests involve two hypotheses:
  - H_0: the null hypothesis, the 'status quo' or current state of belief
  - H_1: the alternative hypothesis, what we're really interested in testing / "proving"
- A hypothesis test allows us to decide whether or not our data provides enough evidence to reject H_0 in favour of H_1, subject to some pre-specified risk of error.
  - hypothesis tests are set up this way so that the burden of proof is on our research hypothesis (H_1) - we carry out our analysis assuming H_0 is true and if the sample data provides strong evidence that it can't possibly be true (subject to some pre-specified level of uncertainty), then we'll feel more "confident" rejecting H_0 (the current state of belief) in favour of H_1 (our research hypothesis)
- The hypotheses are a statistical representation of our research question

1) Hypotheses

- Usually, hypothesis tests involve a parameter (e.g. θ) which characterizes the underlying distribution at the population level
  - ex: a two-sided hypothesis test regarding a parameter θ has the form

    $$\boxed{H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0}$$

    where θ_0 represents some value (number) based on our research question
  - here we are testing whether θ (the "true" parameter value, at the population level) differs "significantly" from θ_0
- Ex: a company claims that an online survey takes 10 minutes to complete. To test whether this is really true, a two-sided hypothesis test can be used. Here the parameter of interest is μ: the average time (at the population level) it takes to complete the online survey. The hypotheses can be written as

  $$H_0 : \mu = 10 \qquad \text{vs.} \qquad H_1 : \mu \neq 10,$$

  This is a two-sided hypothesis test for a population mean μ.

1) Hypotheses

- *Note that we can impose direction in the hypotheses and consider alternatives of the from* $H_1 : \theta > \theta_0$ *or* $H_1 : \theta < \theta_0$*.*
  - In the previous example, if we were interested in testing whether the online survey takes *longer* than 10 minutes on average, our hypotheses could be reformulated as $H_1 : \mu > 10$

2) Test statistic

- A test statistic, T, is essentially a convenient summary of the sample data.
- The form of the test statistic is chosen such that we know its underlying distribution under H_0
  - i.e. if H_0 is true, we know the distribution of T
  - this allows us to determine what values of T are likely if H_0 is true
  - remember: T is a random variable - its value will change from one sample to another, so T has a distribution
- In general, a test statistic has the form

$$\boxed{T = \frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})}}$$

where $\hat{\theta}$ is an estimator of the parameter θ and $\hat{se}(\hat{\theta})$ is an estimator of the standard deviation of $\hat{\theta}$ and θ_0 is the value chosen in the hypotheses. (Note that there are other forms of test statistics.)

2) Test statistic

Note on terminology: estimator vs estimate

An estimator is a rule or formula used to calculate an estimate of some parameter or quantity of interest based on observed data. For example, the sample mean $\bar{X}$ is an estimator of the population mean μ. Once we have observed data we can actually compute the sample mean, that is, we have an estimate - an actual value. In other words, an estimator is the procedure or formula telling us how to use sample data to compute an estimate. An estimator is a random variable since it depends on the sample. The estimate is the actual value obtained once we apply the formula to observed data, it's a number.

2) Test statistic

- Ex: for a hypothesis of the form

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu \neq 0$$

define the test statistic as

$$T = \frac{\bar{X} - 0}{S/\sqrt{n}}$$

where $\bar{X}$ is the sample mean and S is the sample standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

  - Case 1 Normal sample: for a random sample $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, it can be shown that if H_0 is true (i.e. if $\mu = 0$), the test statistic defined above follows a Student t distribution with $n-1$ degrees of freedom.
  - Case 2 large sample: for a large sample size n (usually $n > 30$), the assumption of normality can be relaxed (and the CLT kicks in), and for any *iid* random sample the test statistic $T \approx \mathcal{N}(0,1)$ under H_0.

3) P-value

- P-value: the p-value allows us to decide whether the observed value of the test statistic T is likely under H_0 (i.e. if H_0 is true)
    - the p-value essentially provides a measure of the strength of evidence against H_0 based on the sample
- Formally, the p-value is the probability that the test statistic T (random variable, pre-experiment) is equal to or more extreme than what's observed from the sample data, assuming H_0 is true.
    - recall: the test statistic is chosen in such a way that we know its distribution (and thus plausible values that it can take on) when H_0 is true, so we can compute the p-value!
    - the form of the p-value will depend on the hypotheses being tested, as well as the distribution of the test statistic T

3) P-value

- Calculating the p-value for a two-sided test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

  - suppose that based on a random sample $X_1, \ldots, X_n$ we obtain a test statistic value of T_{obs}
  - for a two-sided test the p-value is $p = P(|T| \geq |T_{obs}| \mid H_0)$, i.e. the probability that the test statistic (in absolute value) is greater than or equal to what's observed (in absolute value), given that H_0 is true
  - usually, the distribution of T is symmetric about 0 so that the p-value simplifies to
$$p = 2 \times P(T \geq |T_{obs}| \mid H_0)$$
  - since we know the distribution of T under H_0, we can compute this probability

3) P-value

Example

- back to the hypotheses

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

e.g.: we're interested in testing the hypothesis that on average, people break even when leaving a casino

- we defined the test statistic as

$$T = \frac{\bar{X} - 0}{S/\sqrt{n}}$$

  - we know that for a normally distributed random sample $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, T follows a Student t distribution with $n-1$ degrees of freedom

- *intuition: if H_0 is true, $\bar{X}$ should be close to 0. We scale the test statistic by $S/\sqrt{n}$ to account for variability. Moreover, the particular form of T is such that we know the distribution under H_0.*

3) P-value

Example (continued)

- Since the Student t distribution is symmetric about 0, the p-value can be computed as
$$p = 2 \times P(t_{n-1} > |T_{obs}|)$$
where t_{n-1} denotes a random variable from the Student t distribution with $n-1$ degrees of freedom
  - to calculate this probability, we can either use a table or a statistical software

3) P-value

One-sided tests:

- for one-sided hypothesis tests, the p-value is computed slightly differently:
  - for $H_1 : \theta > \theta_0$, $p = P(T > T_{obs} | H_0)$
  - for $H_1 : \theta < \theta_0$, $p = P(T < T_{obs} | H_0)$

Rejection region:

- Note that one can use a rejection region rather than a p-value to make a decision.
- The rejection region is determined based on critical values from the distribution of T under H_0.
- If the observed value T_{obs} falls within the rejection region, then we reject H_0 in favour of H_1.

4) Make a conclusion

- Recall that a statistical hypothesis test is a formal procedure that allows us to determine if the data provide enough evidence to reject the null hypothesis H_0 in favour of the alternative H_1
    - i.e. it's a decision rule that, when applied to a sample, allows us to decide if:
        - yes, the data provide enough evidence to reject H_0, in other words accept H_1 (accordingly to some pre-specified risk of error)
        - no, the data do not provide enough evidence to reject H_0, in other words, we cannot accept H_1 beyond reasonable doubt (according to some pre-specified risk of error).
    - $\rightarrow$ there are two possible decisions: reject or do not reject H_0

4) Make a conclusion

- The p-value (or rejection region) allows us to make a decision:
  - If H_0 is indeed true, the p-value should be large, that is, observing the test statistic T_{obs} should be likely if H_0 is really true.
  - If the p-value is small, it means that observing something equal to or more extreme than T_{obs} is very unlikely, and so we're inclined to think that H_0 is not true
- Now there's always some underlying risk that we're making a mistake when we make a decision.
- In statistical tests, there are two types of errors:
  - type I error: we reject H_0 when H_0 is actually true
  - type II error: we fail to reject H_0 when H_0 is actually false

		TRUTH	
		H_0	H_1
DECISION	H_0	✓	type II error
	H_1	type I error	✓

4) Make a conclusion

- Since the distribution of T under H_0 is known, we can control the type I error rate
  - we refer to this as the level of the test, denoted by α

$$\boxed{\alpha = P(\text{ reject } H_0 \mid H_0 \text{ true })}$$

  - i.e. α is the probability that we reject H_0 when H_0 is actually true
- As a researcher, we can fix the level of error risk that we're willing to tolerate, that is, we select the level α
  - often, a value of $\alpha = 0.05$ (5%) is used
- To make a decision, we compare the p-value p with the level of the test α:
  - if $p < \alpha$ we reject H_0
  - if $p > \alpha$ we fail to reject H_0

4) Make a conclusion

- $p < \alpha \rightarrow$ reject H_0 at the α significance level
  - a small p-value suggests that it's very unlikely to observe what we're observing if H_0 is indeed true, which provides evidence against H_0, i.e. H_0 is likely not true
  - the data provide sufficient evidence to reject H_0 in favour of H_1 at the α significance level
- $p > \alpha \rightarrow$ fail to reject H_0 at the α significance level
  - a large p-value suggests that it's very likely to observe what we're observing under H_0, which provides evidence in support of H_0, i.e. H_0 is probably true
  - the data do not provide sufficient evidence to reject H_0 in favour of H_1 at the α significance level

4) Make a conclusion

- It's important to make relevant conclusions in the context of the research problem
  - ex: back to the example $H_0 : \mu = 0$ vs. $H_1 \mu \neq 0$ (on average, people break even at the casino), suppose we obtain a p-value of $p = 0.07$
    - if $\alpha = 0.05$ then $p > \alpha$ and so we fail to reject H_0; the data suggests that on average people do break even at the casino, or equivalently but more convoluted, the data do not provide enough evidence to suggest that people do not break even at the casino (at the $\alpha = 0.05$ significance level)
    - if $\alpha = 0.10$ then $p < \alpha$ and so we reject H_0; the data suggests that on average people do not break even at the casino (at the $\alpha = 0.10$ significance level)
- careful with the interpretation of what a p-value is exactly...
  - the p-value is **NOT** the probability that H_0 is false!
  - the p-value is **NOT** the probability that H_1 is true!
  - the p-value is a **conditional probability statement**, it's the probability of observing something as extreme or even more extreme than T_{obs}, given that H_0 is true

Power of a test

- Recall: in statistical tests, there are two types of errors:
  - type I error: we reject H_0 when H_0 is actually true
  - type II error: we fail to reject H_0 when H_0 is actually false

		TRUTH	
		H_0	H_1
DECISION	H_0	✓	type II error
	H_1	type I error	✓

- Let β represent the probability of a type II error:

$$\beta = P(\text{ fail to reject } H_0 \mid H_0 \text{ is false })$$

- The power of a test is the probability that the test will reject H_0 when H_0 is actually false, thus

$$\text{power} = 1 - \beta = P(\text{ reject } H_0 \mid H_0 \text{ is false })$$

Tests about μ

- Often, we're interested in carrying out hypothesis tests involving the population mean μ
- The t-test is a well known hypothesis test for the population mean
- There are several versions of t-tests depending on the context of the problem:
    - one sample t-test
    - independent two sample t-test
    - paired samples t-test

T-test

- **one sample t-test:** involves a single population mean μ
  - ex: $H_0 : \mu = 50000$ vs. $H_1 : \mu \neq 50000$ where μ represents the mean income in Quebec
- **independent two sample t-test:** involves two population means μ_1 and μ_2 from two independent populations
  - ex: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ where μ_1 and μ_2 respectively represent the mean salary in Quebec and Ontario
  - we will revisit this test in a regression context
- **paired samples t-test:** involves two population means μ_1 and μ_2 from dependent (or paired) populations
  - ex: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ where μ_1 and μ_2 respectively represent the mean annual health spending of residents in Quebec before and after a reform on the health system; we would have measurements on spending for each subject in the sample both before and after the reform

Paired t-test

- In a paired t-test, the problem of comparing two population means can be converted back into a test for a single population mean!
  - instead of analyzing each group separately, we can consider the paired differences between the two groups
  - e.g. rather than comparing spending before X_{before} and after the reform X_{after}, consider the difference in spending for each subject $X_D = X_{after} - X_{before}$. The hypotheses then go from

    $$H_0 : \mu_{after} = \mu_{before} \quad \text{vs.} \quad H_1 : \mu_{after} \neq \mu_{before}$$

    to

    $$H_0 : \mu_D = 0 \quad \text{vs.} \quad H_1 : \mu_D \neq 0$$

- One advantage of this type of design is that each subject acts as their own control
  - other variables that could affect the results (age, income, education level, etc.) are automatically controlled for

Hypothesis testing: example

- It's well known that it's dangerous to text while driving, but is it dangerous to text and walk?
- That's the question researchers at Tech3Lab wanted to investigate.
    - Tech3Lab (http://tech3lab.hec.ca/) is an applied research laboratory in management sciences that specializes in the analysis of interactions between business technologies and employees or consumers.

Le risque de texter en marchant :
La lenteur du désengagement
OBJECTIF : COMPRENDRE LES PROCESSUS COGNITIFS SOUS-JACENTS AU MULTITÂCHE MOBILE
HYPOTHÈSE
L'engagement cognitif ralentit la capacité attentionnelle du piéton
MESURES
Électroencéphalographie
MOMENT D'ENGAGEMENT
MOMENT DE DÉSANGAGEMENT
MOMENT D'ENGAGEMENT
Avec la permission de ICI Radio-Canada

Hypothesis testing: example

Study set-up:

- $n = 35$ subjects participated in the study.
- Each person had to walk on a treadmill in front of a screen where obstacles were projected.
- In one of the sessions, the subjects walked while talking on a cell phone, whereas in another session, they walked while texting.
  - The order of these sessions was determined at random. *Why do you think this is important?*
- Different obstacles were randomly projected during the session.
  - We are only interested in one kind of projection: a cyclist riding towards the participant.

Hypothesis testing: example

- Research question: Is the reaction time different when texting on a phone in comparison to talking?
- Variable of interest: time (in seconds) that it takes for a person to notice the obstacle when walking while texting vs. talking on a cell phone.
- Data: each subject is submitted to both experimental conditions (texting/walking and talking/walking) → we have paired samples

Subject	Time (talking)	Time (texting)
1	1.487	1.601
2	1.385	2.063
3	0.459	1.947
4	0.797	1.034
⋮	⋮	⋮

Like most of the examples seen in this course, the data were generated randomly and the results are fictitious.

Hypothesis testing: example

- Population: adults 18 or older
- Sample: $n = 35$ subjects
- Variables measured:
  - time to perceive the obstacle (quantitative)
  - type of distraction (talking or texting) (qualitative, nominal)
- Type of statistical analysis: paired t-test
  - we will test whether the *difference* in reaction time between talking and texting is zero

Hypothesis testing: example

- Notation:
  - $C_i =$ the reaction time for talking while walking for subject i
  - $T_i =$ the reaction time for texting while walking for subject i
  - $D_i = T_i - C_i =$ difference in reaction time between texting and talking for subject i
  - μ_T : population level mean reaction time for texting while walking
  - μ_C : population level mean reaction time for talking while walking
  - μ_D : population level mean difference in reaction time between texting and talking
- the research problem can be written in terms of the following hypotheses
$$H_0 : \mu_T = \mu_C \quad \text{vs.} \quad H_1 : \mu_T \neq \mu_C$$
or, equivalently as
$$H_0 : \mu_D = 0 \quad \text{vs.} \quad \mu_D \neq 0$$

Hypothesis testing: example

Calculations:

- $n = 35$
- $\bar{D} = 0.313$
  - Why can't we simply say $\bar{D} \neq 0$ so we can reject H_0? Recall: $\bar{D}$ is a (continuous) random variable, and in fact $P(\bar{D} = 0) = 0$! We look at a test statistic, rather, to evaluate whether $\bar{D}$ is sufficiently far from 0 to make a decision. We must account for the variability in $\bar{D}$.
- $S = 0.6369$
- the test statistic is

$$T = \frac{\bar{D} - 0}{S/\sqrt{n}} = 2.91$$

- recall: under H_0, $T \sim t_{n-1}$, i.e. T follows a Student t distribution with 34 degrees of freedom → this allows to compute the p-value:

$$p = 2 \times P(t_{34} > 2.91) = 0.0064$$

all of these results are obtained from R using the t.test function, see the sample code

Hypothesis testing: example

Conclusion:

- $p = 0.0064 < \alpha = 0.05 \Rightarrow$ we reject H_0 and can conclude that there is sufficient evidence (at the $\alpha = 5\%$ level) that there is a significant difference between the reaction times of people who are texting vs. talking
- Can we conclude that the reaction time for those texting is longer than those talking?

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Confidence Intervals

- A confidence interval is a convenient way to report results as

$$\boxed{\text{point estimate} \;\pm\; \text{margin of error}}$$

  - it provides a point estimate along with some margin of error
  - it thus gives some indication of the variability inherent in the estimation procedure

- More formally, a $100(1-\alpha)\%$ confidence interval for a parameter θ has the general form

$$\hat{\theta} \pm Q_{\alpha/2}\hat{se}(\hat{\theta})$$

where

  - $\hat{\theta}$ is an estimator of the parameter θ
  - $\hat{se}(\hat{\theta})$ is an estimator of the standard deviation of $\hat{\theta}$
  - $Q_{\alpha/2}$ is the $1-\alpha/2$ quantile from the distribution of the statistic $(\hat{\theta}-\theta)/\hat{se}(\hat{\theta})$

Confidence intervals

$$\boxed{(\hat{\theta} - Q_{\alpha/2}\hat{se}(\hat{\theta}),\ \hat{\theta} + Q_{\alpha/2}\hat{se}(\hat{\theta}))}$$

- Note: the bounds of the C.I. are random variables!
  - this follows since both $\hat{\theta}$ and $\hat{se}(\hat{\theta})$ are random variables - their values depend on the sample and will vary from one sample to another
- Ex: for a random sample $X_1, \ldots, X_n$ from a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, it can be shown that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

i.e. it follows a Student t distribution with $n-1$ degrees of freedom. Let $t_{n-1,\alpha/2}$ denote the $1-\alpha/2$ quantile from a Student t distribution with $n-1$ degrees of freedom. Then a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}}$$

Confidence intervals

Interpretation

- Careful with interpreting confidence intervals - this is a notion that students often get confused with
- before actually calculating the C.I., there is a $100(1-\alpha)\%$ chance that θ is contained in the random interval

$$(\hat{\theta} - Q_{\alpha/2}\hat{se}(\hat{\theta}),\ \hat{\theta} + Q_{\alpha/2}\hat{se}(\hat{\theta}))$$

- after we obtain a sample and actually compute the C.I., there is no more notion of probability!
    - the true value of the parameter θ is either in the compute C.I. or not, there is no more probability as everything takes on a numerical value
- interpretation: if we were to repeat the experiment many many many times, and calculate a $100(1-\alpha)\%$ C.I. each time, then roughly $100(1-\alpha)\%$ of the C.I.s would contain the true value of θ

Confidence Intervals

- When it comes to C.I., our *confidence* is in the procedure and not the results:
  - the recipe is good:
    - i.e. the procedure used to compute C.I.s has a $100(1-\alpha)\%$ chance of containing the true parameter θ; and we get to choose the level α
  - the cake either turns out good or bad:
    - i.e. the numerical results obtained from the sample either contains θ or not (although, unlike a cake where we can simply taste it, we never know if the true θ is in the computed C.I. or not)

Confidence intervals: example

Back to our example about texting vs. talking while walking

- Recall: we were interested in μ_D: the mean difference in reaction time between texting and talking while walking
- We obtained a point estimate $\hat{\mu}_D = \bar{X} = 0.313$ and standard error $\hat{se}(\hat{\mu}_D) = S/\sqrt{n} = 0.1077$
- We use the quantiles from the Student t distribution with $n - 1 = 34$ degrees of freedom
  - using $\alpha = 0.05$, we have that $t_{34,0.025} = 2.03$
- a 95% C.I. for μ_D is then

$$0.313 \pm 2.03 \times 0.1077 = (0.094, 0.532)$$

there may be slight difference in the exact numerical answers due to rounding

- **note:** most statistical softwares (including R) will provide estimates, standard errors, confidence intervals, p-values, etc. by default when doing the analysis; no need to compute them “by hand”!

Confidence intervals: example

- we obtained a 95% C.I. $(0.094, 0.532)$, what does this mean?
  - it does NOT mean that there is a 95% probability that the population mean μ_D is between 0.094 and 0.532
  - remember μ_D, although unknown, is a fixed value that characterizes the population, i.e. the underlying distribution of random variable D
  - thus μ_D is either in the interval $(0.094, 0.532)$ or not, there is no probabilistic statement to make here
- it does mean that: if the study were redone under the exact same conditions over and over and over again (a large number of times) and a C.I. was calculated each time in this same way, we would expect 95% of these intervals to contain the true value of μ_D

Confidence intervals and hypothesis testing

- There is a parallel between (two-sided) hypothesis tests and confidence intervals: both approaches can be used to test hypotheses and will lead to the same conclusion
  - for hypotheses
    $$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$
    the following are equivalent:
    - construct a $100(1-\alpha)\%$ C.I. for θ, if θ_0 is not in the interval then we can reject H_0 at the α significance level
    - carry out a hypothesis test at level α, if the p-value $p < \alpha$ then we can reject H_0 at the α significance level
- Ex: in our example about texting vs. talking while walking, we were interesting in testing $H_0 : \mu_D = 0$ vs. $H_1 : \mu_D \neq 0$
  - we obtained a p-value of $p = 0.0064$; with $\alpha = 0.05$, we have that $p < \alpha$ so reject H_0 at the $\alpha = 5\%$ level
  - we obtained a 95% C.I. for μ_D of $(0.094, 0.532)$, since the C.I. does not contain 0 we can reject H_0 at the $\alpha = 5\%$ level

Confidence intervals and hypothesis testing

More on confidence intervals...

- Let's focus on our previous example about texting vs. talking while walking
  - Recall: D_i : represents the difference in reaction time for subject i when texting vs. talking
- Assuming $D_i \sim \mathcal{N}(\mu, \sigma^2)$, we have that

$$\bar{D} \sim \mathcal{N}(\mu, \sigma^2/n), \quad \text{or, equivalently} \quad \frac{\bar{D}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

- Recall: for $Z \sim \mathcal{N}(0,1)$, the 97.5% quantile is 1.96, in other words, $P(-1.96 \leq Z \leq 1.96) = 0.95$
  - (i.e. for observations coming from a standard normal distribution, roughly 95% will fall between ± 1.96)
- Thus

$$P\left(-1.96 \leq \frac{\bar{D}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

which can be rearranged to

$$P\left(\bar{D} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{D} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Confidence intervals and hypothesis testing

- This is exactly the definition of a 95% C.I. for μ: there is a 95% probability that μ falls within the RANDOM interval $\bar{D} \pm 1.96\frac{\sigma}{\sqrt{n}}$, thus a 95% C.I. for μ is

$$\left(\bar{D} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{D} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

  - Careful: don't forget, the bounds of above C.I. are random variables since $\bar{D}$ is a random variable. When we replace the bounds by the actual observed sample values, we no longer have a probabilistic statement.

- This is not exactly the form for C.I.s that we saw earlier...
  - in practice, we rarely know the true value of the population variance σ, and we thus estimate it by the sample variance S
  - in doing so, we change the underlying distribution to that of Student t with $n-1$ degrees of freedom:

$$\frac{\bar{D} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Confidence intervals and hypothesis testing

- Working with this, we have

$$P\left(-t_{n-1,0.025} \leq \frac{\bar{D}-\mu}{S/\sqrt{n}} \leq t_{n-1,0.025}\right) = 0.95$$

which can be rearranged in to give a 95% C.I.

$$\left(\bar{D} - t_{n-1,0.025}\frac{S}{\sqrt{n}}, \bar{D} + t_{n-1,0.025}\frac{S}{\sqrt{n}}\right)$$

Confidence intervals and hypothesis testing

How are C.I.s and hypothesis tests equivalent?

- Focusing on our example again, with $\alpha = 0.05$, $H_0 : \mu_D = 0$ and $H_1 : \mu_D \neq 0$
- The bounds of the C.I. are computed according to

$$\bar{D} \pm t_{n-1,0.025} \frac{S}{\sqrt{n}}$$

- If the C.I. does NOT contain 0 it means that either:

i) both the lower and upper bounds are greater than 0 and thus

$$\bar{D} - t_{n-1,0.025} \frac{S}{\sqrt{n}} > 0 \quad \text{i.e.} \quad \frac{\bar{D}}{S/\sqrt{n}} > t_{n-1,0.025}$$

ii) OR both the lower and upper bounds are lower than 0 and thus

$$\bar{D} + t_{n-1,0.025} \frac{S}{\sqrt{n}} < 0 \quad \text{i.e.} \quad -\frac{\bar{D}}{S/\sqrt{n}} > t_{n-1,0.025}$$

* *since the quantity $\bar{D} + t_{n-1,0.025} \frac{S}{\sqrt{n}}$ is negative*

Confidence intervals and hypothesis testing

How are C.I.s and hypothesis tests equivalent?

- An equivalent way to express both inequalities together is

$$\left|\frac{\bar{D}}{S/\sqrt{n}}\right| > t_{n-1,0.025}$$

  - *Note: this is a* ***rejection region*** *- we reject H_0 whenever the value of the test statistic $T = \frac{\bar{D}}{S/\sqrt{n}}$ is greater (in absolute value) than the critical value $t_{n-1,0.025}$*

- In our example, we observed $T_{obs} = 2.91$ and $t_{n-1,0.025} = t_{34,0.025} = 2.03$, and so $2.91 > 2.03$, i.e. $T > t_{n-1,0.025}$
  - thus we must have that $P(T > 2.91) < 0.025$
  - and thus

$$P(|T| > 2.91) = 2 \times P(T > 2.91) < 2 \times 0.025 = 0.05$$

    i.e. the p-value $p < 0.05$ and so we reject H_0

- Thus, if 0 is not in the 95% C.I. for μ, it's equivalent to having a p-value $p < 0.05$

Table of Contents

1 Review: Probability

2 Review: Random Variables
- Discrete Distributions
- Continuous Distributions
- Moments

3 Review: Inference
- Basic Notions
- Sampling and Estimation
- Hypothesis Tests
- Confidence Intervals

4 Concluding remarks

Concluding remarks

- When carrying out a hypothesis test, be sure to:
  - formally write down the hypotheses
  - define all variables / parameters
    - e.g.: don't simply write $H_0 : \mu = 0$, define μ in the context of the problem
  - make a conclusion in the context of the problem
    - e.g.: don't simply say "we reject H_0", say "we can reject H_0 at the α significance level since $p < \alpha$ and we can thus conclude that there is a significant difference in the reaction times between those who text while walking vs. those who talk while walking"
- Always show all relevant work (or show relevant output from R)
- Careful with interpretations, especially for p-values and C.I.s

Concluding remarks

- We reviewed A LOT of material here, DON'T PANIC!
  - it's OK if you don't remember everything we went over, as long as it seems vaguely familiar
  - you will not be tested on any of this review material directly in the final exam
  - this material is, however, very important as it serves as a base for what we will be learning, and some concepts will also be revisited within a regression framework
    - ex: we will be using hypothesis testing and confidence intervals throughout the semester within a regression framework
    - ex: we will revisit t-tests in the context of linear regression models
    - ex: we will learn about logistic (Bernoulli) regression models and Poisson regression models
  - although we won't rigorously go over all of the mathematical details behind the models we learn, know that much of the statistical results are based on concepts we reviewed here (e.g. the CLT is an essential result used in so many statistical "proofs")