Lecture notes on statistics, a basic course

Andersen Ang @ ECS, University of Southampton, UK
andersen.ang@soton.ac.uk Homepage: angms.science First draft: September 1, 2023 version: December 1, 2023

- These lecture notes are short summary of key points, not a substitution of a textbook
- These lecture notes focus on mathematical statistics at the level of undergraduate course for engineering/science degree
  - The notes are not about manipulation of data at the level of business statistics
  - The notes are not at the level of measure-theoretic statistics in pure mathematics

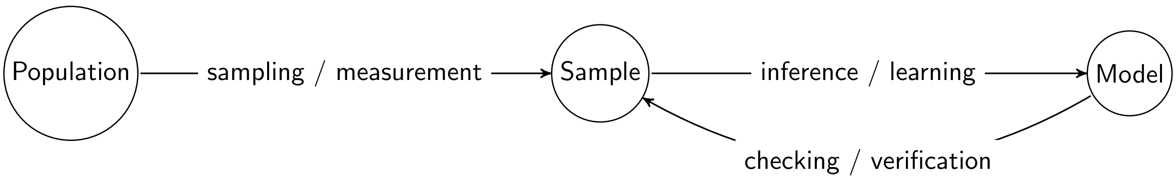
Prerequisite: sufficient knowledge of naive set theory and single variable calculus.

For COMP1215: Ch1 - Ch4, Ch 6.2.1 (only the formula of unbiased estimator of variance), Ch7.1-Ch7.2, Ch 10 and Ch 11.

Contents

1 What is data science

- Data science process



  Three elements

  - **Population**: collection of objects
  - **Sample**: subset of population
  - **Model**: a description of the population learned from the sample

  Three operations

  - **sampling**: select a subset of (finite) object (at random) from population
  - **inference**: fitting a model to a sample
  - **checking**: examining the goodness-of-fit, compatibility of a model to the sample

- Type of data
  - Categorical
    e.g., sex, country of birth
  - Numeric-discrete: $\mathbb{N}$
    e.g., number of phones, number of days (in whole number)
  - Numeric-continuous: $\mathbb{R}$
    e.g., weight, height

- Example of model
  - Statistical distribution: describe observed data $\{x_1, x_2, \ldots, x_n\}$ by $p(x \,|\, \theta)$.
  - Classifier: given $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where x is data attribute and y is class label, find a classifier that give a new data x a label y
  - Regression
  - Clustering: given $\{x_1, x_2, \ldots, x_n\}$, divid them into subsets

- Random variable

- we use the language of probability to describe data
- we treat them as realisations of random variables
- why probability / random variable
  * randomness due to measurement error
  * randomness due to unmeasured factors
  * randomness due to sampling

2 Random variable

- **Sample space** := the set of possible value
- A random variable (RV) is a variable that takes on a value from a sample space with specific probabilities
- **Example** Let $\mathcal{X} = \{1, 2, 3\}$, let X be a random variable over $\mathcal{X}$ with

$$\text{(Split-function expression of RV)} \; : \; X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases} \tag{1}$$

  - $\mathcal{X} = \{1, 2, 3\}$ is the sample space: the possible outcome of X is $1, 2, 3$
  - X is the symbol that denotes the random variable here
  - $X = x$ for a particular value x is called a *realisation* of a random variable
  - A sample is a *realisation* of a random variable
  - Realisation = a fancier term for *observed value*
  - Example of a 2-sample realisation: $\{3, 3\}$
  - Example of a 4-sample realisation: $\{3, 3, 1, 2\}$
- **Probability distribution** $\mathbb{P}(X = x), x \in \mathcal{X}$ is the mathematical notation of random variable, it means

  the probability that the RV X takes the value x in $\mathcal{X}$.

  Using $\mathbb{P}(X = x)$, the split expression (1) can be expressed as a function

$$\text{(Probability expression of RV)} \; : \; \begin{array}{l} \mathbb{P}(X = 1) = 1/2 \\ \mathbb{P}(X = 2) = 1/4 \\ \mathbb{P}(X = 3) = 1/4 \end{array}$$

- **Compact shorthand** $p(x) := \mathbb{P}(X = x), x \in \mathcal{X}$.

$$\text{(Compact expression of RV)} \; : \; \begin{array}{l} p(1) = 1/2 \\ p(2) = 1/4 \\ p(3) = 1/4 \end{array}$$

- **Axiom of probability**
  1. $p(x) \geq 0$, probability cannot be negative
  2. $p(\mathcal{X}) = 1$, the sample space has probability 1
  3. The probability of $X \in A_1$ or $X \in A_2$ with $A_1, A_2 \subset \mathcal{X}$ is

$$\mathbb{P}(X \in A_1 \cup A_2) \;=\; \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2) - \mathbb{P}(X \in A_1 \cap A_2) \tag{inclusion-exclusion principle}$$

  and if $A_1 \cap A_2 = \varnothing$, then

$$\mathbb{P}(X \in A_1 \cup A_2) \;=\; \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2) \tag{σ-additivity}$$

  Note: saying $\boxed{p(x) \leq 1 \text{ is a probability axiom}}$ is wrong. It can be derived from the 3 axioms above.

- **Example** Given $\mathbb{P}(X = 1) = a, \mathbb{P}(X = 2) = b$ and $\mathbb{P}(X = 3) = c$, find $\mathbb{P}(X \geq 2)$

$$\begin{aligned} \mathbb{P}(X \geq 2) &= \mathbb{P}(X \in \{2\} \cup \{3\}) \\ &= \mathbb{P}(X \in \{2\}) + \mathbb{P}(X \in \{3\}) - \mathbb{P}(X \in \{2\} \cap \{3\}) \\ &= b + c - 0 = b + c. \end{aligned}$$

- **Product rule** $\mathbb{P}(X \in A_1 \text{ and } X \in A_2) = \mathbb{P}(X \in A_1)\mathbb{P}(X \in A_2)$
- **Complement rule** $\mathbb{P}(X < x) = 1 - \mathbb{P}(X \geq x)$.
  - This is useful for continuous RV.
  - Pay attention to the equality sign, it is $X < x$, not $X \leq x$
  - Similarly we have $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$.
- **Example** $\mathbb{P}(a < X < b) = \mathbb{P}(X > a \text{ AND } X < b) = \mathbb{P}(X > a)\mathbb{P}(X < b) = \Big(1 - \mathbb{P}(X \leq a)\Big)\Big(1 - \mathbb{P}(X \geq b)\Big)$
- **Two RVs**
  - We now consider (X, Y) for two RVs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
    - We now have two sample spaces: $\mathcal{X}$ and $\mathcal{Y}$
    - We have two RVs: X and Y
  - Important: (X, Y) is not the same as $X + Y$
    - $(X, Y) \in \mathcal{X} \times \mathcal{Y}$
      - (X, Y) is an ordered pair, it has two numbers
      - $\mathcal{X} \times \mathcal{Y}$ is the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$
    - $X + Y \in \mathcal{X} \oplus \mathcal{Y}$
      - $X + Y$ gives a single number in the end, it is not a pair
      - $\mathcal{X} \oplus \mathcal{Y}$ is the Minkowski sum of $\mathcal{X}$ and $\mathcal{Y}$
  - **Example** Toss a 2-sided dice X and a 4-sided dice Y, we have

$$\mathcal{X} \times \mathcal{Y} = \{1,2\} \times \{1,2,3,4\} = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4)\} = \text{all possible outcome-pairs}$$

$$\mathcal{X} \oplus \mathcal{Y} = \{1,2\} \oplus \{1,2,3,4\} = \{2,3,4,5,6\} = \text{all possible sum}$$

- **The probability distribution of two RVs as a table** For example

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	0.05	0.15	0.1
$Y = 2$	0.25	0.15	0.3

  - Each box represents a particular **joint probability** of $X = x$ and $Y = y$, denoted as $\mathbb{P}(X = x, Y = y)$, which means the probability of $\boxed{X = x \text{ AND } Y = y}$.
  - Example

$$\mathbb{P}(X = 1, Y = 1) = 0.05$$
$$\mathbb{P}(X = 1, Y = 2) = 0.25$$
$$\mathbb{P}(X = 2, Y = 1) = 0.15$$

  - Recall axiom of probability: probability of sample space is 1, so **the sum of all boxes must be 1.** So **always normalize the table such that the sum of all boxes is 1.** Mathematically

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) = 1.$$

- **Sum rule and marginal probability**

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y). \qquad \text{(marginal probability)}$$

  It means the probability of $\boxed{X = x \text{ ignoring } Y}$.

- We get marginal probability from the table

	$X = 1$	$X = 2$	$X = 3$	
$Y = 1$	0.05	0.15	0.1	$0.3 = \mathbb{P}(Y = 1)$
$Y = 2$	0.25	0.15	0.3	$0.7 = \mathbb{P}(Y = 2)$
	$0.3 = \mathbb{P}(X = 1)$	$0.3 = \mathbb{P}(X = 2)$	$0.4 = \mathbb{P}(X = 3)$	

Make sure all probabilities have to follow the 3 axioms.

- The sum of all joint probability must be 1
- The sum of all marginal probability on X must be 1
- The sum of all marginal probability on Y must be 1
- All probabilities ≥ 0
- Inclusion-exclusion principle holds for any combinations

If any thing above is violated, that means you are wrong.

- **Conditional probability**

$$\mathbb{P}(X=x|Y=y) \;:=\; \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(Y=y)} \qquad \text{conditional} = \frac{\text{joint}}{\text{marginal}} \qquad \text{(Conditional probability)}$$

It means the probability of $\boxed{X=x \textbf{ given } Y=y}$

  - $\mathbb{P}(X=x|Y=y)$ and $\mathbb{P}(X=x)$ are two different things

- **Example** $\mathcal{X}=\{1,2,3\}, \mathcal{Y}=\{1,2\}$

	$X=1$	$X=2$	$X=3$	
$Y=1$	0.05	0.15	0.1	$0.3=\mathbb{P}(Y=1)$
$Y=2$	0.25	0.15	0.3	$0.7=\mathbb{P}(Y=2)$
	$0.3=\mathbb{P}(X=1)$	$0.3=\mathbb{P}(X=2)$	$0.4=\mathbb{P}(X=3)$	

We have

$$\mathbb{P}(X=1|Y=1)=\mathbb{P}(X=1,Y=1)/\mathbb{P}(Y=1)=0.05/0.3=1/6.$$

$$\mathbb{P}(X=1|Y=2)=\mathbb{P}(X=1,Y=2)/\mathbb{P}(Y=2)=0.25/0.7=5/14.$$

- **Independent RV** If $\mathbb{P}(X=x,Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y)$ for all $x\in\mathcal{X}, y\in\mathcal{Y}$ then X,Y are independent
  - This also implies $\mathbb{P}(X=x|Y=y)=\mathbb{P}(X=x)$
    X,Y are independent, knowing Y tells nothing about X

- **Independent and identically distributed (i.i.d.)** X,Y are i.i.d. if
  1. independent $\mathbb{P}(X=x,Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y)$
  2. identically distributed $\mathbb{P}(X=x)$, $\mathbb{P}(Y=y)$ follow the same *probability distribution function*.

- **Discrete vs continuous random variable**

Discrete RV

- $\mathcal{X}$ is a finite set, discrete set
- we call distribution a probability mass function (PMF)
- Axiom of probability on x is nonnegative

$$p(x)\geq 0 \;\forall x\in\mathcal{X}$$

- Axiom of probability on sample space

$$p(\mathcal{X}) := \sum_{x\in\mathcal{X}} p(x) = 1$$

- Axiom of interval

$$\mathbb{P}(a\leq X\leq b)=\sum_a^b p(x)$$

- $\mathbb{P}(X=x)$ can be 0 or not 0
  * $\mathbb{P}(X\leq a)=\mathbb{P}(X<a)+\mathbb{P}(X=a)$

Continuous RV

- $\mathcal{X}$ is an infinite set, an interval
- we call distribution a probability density function (PDF)
- Axiom of probability on x is nonnegative

$$p(x)\geq 0 \;\forall x\in\mathcal{X}$$

- Axiom of probability on sample space

$$p(\mathcal{X}) := \int_{\mathcal{X}} p(x)dx = 1$$

- Axiom of interval

$$\mathbb{P}(a\leq X\leq b)=\int_a^b p(x)dx$$

- $\mathbb{P}(X=x)$ is always 0, this confusing result is from real analysis (advanced calculus)
  * $\mathbb{P}(X\leq a)=\mathbb{P}(X<a)$

- **Cumulative distribution function**

$$\mathbb{P}(X \leq x) = \begin{cases} \displaystyle\sum_{t \leq x} p(x) & \text{(discrete RV)} \\ \displaystyle\int_{-\infty}^{x} p(t)dt & \text{(continuous RV)} \end{cases}$$

- **Quantile function / inverse CDF**

$$Q(p) = \Big\{ x \in \mathcal{X} \ : \ \mathbb{P}(X \leq x) = p \Big\}$$

It means the value x such that $\mathbb{P}(X \leq x)$ is p

  - Median is defined as $p = 1/2$, i.e.,

$$Q(0.5) = \Big\{ x \in \mathcal{X} \ : \ \mathbb{P}(X \leq x) = 0.5 \Big\}$$

3 Expectation and variance

- Notation: we write $\mathbb{P}(X = x)$ compactly as $p(x)$
- **Expected value** of a RV is $\mathbb{E}[X]$

$$\mathbb{E}[X] \coloneqq \begin{cases} \displaystyle\sum_{x \in \mathcal{X}} xp(x) & \text{discrete RV} \\ \displaystyle\int_{\mathcal{X}} xp(x)dx & \text{continuous RV} \end{cases}$$

For discrete RV in table form: if X is a RV with the distribution

x	x_1	x_2	$\cdots$	x_n
$\mathbb{P}(X = x)$	p_1	p_2	$\cdots$	p_t

then $\mathbb{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$

In other words, expectation = weighted sum

  - the weights are $p(x)$, interpreted as "occurrence frequency"

The other name of expected value is mean

- **Sample mean** of an observed dataset $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ is

$$\overline{x} \coloneqq \frac{1}{n} \sum_{i=1}^{n} x_i.$$

  - Expected value $\neq$ sample mean
  - We may never know what is the exact value of $\mathbb{E}[X]$
  - We are *using sample mean to estimate the population expected value*
  - Sample mean is depending on the data we obtain, while population mean does not

- **Expected value** of a function of a RV is $\mathbb{E}[f(X)]$

$$\mathbb{E}[f(X)] \coloneqq \begin{cases} \displaystyle\sum_{x \in \mathcal{X}} f(x)p(x) & \text{discrete RV} \\ \displaystyle\int_{\mathcal{X}} f(x)p(x)dx & \text{continuous RV} \end{cases}$$

- **Variance**: When $f(\cdot) = (\ \cdot - \mathbb{E}[\ \cdot\ ])^2$, we have the **variance**

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \begin{cases} \displaystyle\sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x) & \text{discrete RV} \\ \displaystyle\int_{\mathcal{X}} (x - \mathbb{E}[X])^2 p(x)dx & \text{continuous RV} \end{cases}$$

For discrete RV in table form: if X is a RV with the distribution

x	x_1	x_2	$\cdots$	x_n
$\mathbb{P}(X = x)$	p_1	p_2	$\cdots$	p_t

then $\mathbb{V}[X] = (x_1 - \overline{x})^2 p_1 + (x_2 - \overline{x})^2 p_2 + \cdots + (x_n - \overline{x})^2 p_n$

- **Example** For $\mathcal{X} = \{1,2,3\}$ with $p(1) = 0.5, p(2) = 0.4$ and $p(3) = 0.1$, let $f(x) = x^2$, then

$$\begin{aligned}
\mathbb{E}[X] &= 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 &&= 1.6 \\
\mathbb{E}[f(X)] &= 1^2 \cdot 0.5 + 2^2 \cdot 0.4 + 3^2 \cdot 0.1 &&\approx 3 \\
\mathbb{V}[X] &= (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 &&= 0.44
\end{aligned}$$

- **Standard deviation** is $\sqrt{\mathbb{V}[X]}$
  - Variance is denoted as σ^2
  - Standard deviation is denoted as σ
  - Why we have variance and standard deviation: mean has unit^1, variance has the unit^2, to make variance comparable to mean, we take squared-root to get standard deviation, with the unit^1
- $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, useful
  From the above example, $\mathbb{E}[X^2] = 3$ and $\mathbb{E}[X] = 1.6$, we have $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 3 - 1.6^2 = 0.44$
- Some non-trivial facts
  - If X is a discrete RV, it is possible that $\mathbb{E}[X] \notin \mathcal{X}$
    * Example is coin flip: $\mathcal{X} = \{0,1\}$, but the expected value of a fair coin is $\dfrac{1}{2} \cdot 0 + \dfrac{1}{2} \cdot 1 = 0.5 \notin \mathcal{X}$
  - For some RV, expectation does not exist. E.g., for Cauchy distribution, expectation is undefined.
  - For some RV, expectation exists, but it is infinite
  - Same for variance: it can be undefined or infinite.
- **Sample variance / Unbiased estimator of variance** of an observed dataset $\boldsymbol{x} = (x_1, ..., x_n)$ is

$$s_{\boldsymbol{x}}^2 \coloneqq \frac{1}{n-1} \sum_{i=1}^{n} \big(x_i - \overline{x}\big)^2.$$

  - $s_{\boldsymbol{x}}^2$ means "the sample variance from observed data $\boldsymbol{x}$"
  - Note that we are dividing by $n-1$, NOT n
  - The sample standard deviation $s_{\boldsymbol{x}}$ is just the squared-root of $s_{\boldsymbol{x}}^2$

3.1 Advanced topics on expectation and variance

- **Expectation of function of two RVs**

$$\mathbb{E}[f(X,Y)] = \begin{cases} \displaystyle\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y)p(x,y) & \text{discrete RV} \\ \displaystyle\int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y)p(x,y)dxdy & \text{continuous RV} \end{cases}$$

- **Example** Suppose $\mathcal{X} = \{1,2,3\}, \mathcal{Y} = \{1,2\}$

	$X=1$	$X=2$	$X=3$	
$Y=1$	0.05	0.15	0.1	$0.3 = \mathbb{P}(Y=1)$
$Y=2$	0.25	0.15	0.3	$0.7 = \mathbb{P}(Y=2)$
	$0.3 = \mathbb{P}(X=1)$	$0.3 = \mathbb{P}(X=2)$	$0.4 = \mathbb{P}(X=3)$	

  Then
  - For $f(x,y) = xy$, we have $\mathbb{E}[XY] = \displaystyle\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} xyp(x,y)$

$$\begin{aligned}
\mathbb{E}[XY] &= (1\times 1)0.05 + (1\times 2)0.25 + (2\times 1)0.15 + (2\times 2)0.15 + (3\times 1)0.1 + (3\times 2)0.3 \\
&= 0.05 + 0.5 + 0.3 + 0.6 + 0.3 + 1.8 \\
&= 3.55
\end{aligned}$$

– For $f(x,y)=x+y$, we have $\mathbb{E}[X+Y]=\displaystyle\sum_{x\in\mathcal{X},y\in\mathcal{Y}}(x+y)p(x,y)$

$$\begin{aligned}\mathbb{E}[X+Y] &= (1+1)0.05+(1+2)0.25+(2+1)0.15+(2+2)0.15+(3+1)0.1+(3+2)0.3\\ &= 0.1+0.75+0.45+0.6+0.4+1.5\\ &= 3.8\end{aligned}$$

– For $f(x,y)=(x,y)$, we have $\mathbb{E}[(X,Y)]=\displaystyle\sum_{x\in\mathcal{X},y\in\mathcal{Y}}(x,y)p(x,y)$

$$\begin{aligned}\mathbb{E}[(X,Y)] &= (1,1)0.05+(1,2)0.25+(2,1)0.15+(2,2)0.15+(3,1)0.1+(3,2)0.3\\ &= (0.05,0.05)+(0.25,0.5)+(0.3,0.15)+(0.3,0.3)+(0.3,0.1)+(0.9,0.6)\\ &= (2.1,1.7)\end{aligned}$$

- **Expectation is linear**
$$\mathbb{E}[f(X)+g(Y)]=\mathbb{E}[f(X)]+\mathbb{E}[g(Y)].$$
This implies the following useful equality: for any a,b,c,
$$\mathbb{E}[aX+bY+c]=a\mathbb{E}[X]+b\mathbb{E}[Y]+c.$$

- Expectation of *independent* RVs
$$\mathbb{E}[f(X)g(Y)]=\mathbb{E}[f(X)]\cdot\mathbb{E}[g(Y)].$$
Proof
$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} f(x)g(y)p(x,y)\\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} f(x)g(y)p(x)p(y) \quad X,Y\text{independent so } p(x,y)=p(x)p(y)\\ &= \sum_{x\in\mathcal{X}} f(x)p(x)\sum_{y\in\mathcal{Y}} g(y)p(y)\\ &= \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\end{aligned}$$

- **Variance quadratic formula**
$$\mathbb{V}[aX\pm bY+c]=a^2\mathbb{V}[X]\pm 2ab\,\mathrm{cov}(X,Y)+b^2\mathbb{V}[Y].$$
This implies that if X and Y are independent (so $\mathrm{cov}(X,Y)=0$)
$$\mathbb{V}[aX\pm bY+c]=a^2\mathbb{V}[X]+b^2\mathbb{V}[Y].$$

- **Taylor series approximation** We want to find $\mathbb{E}[f(X)],\mathbb{V}[f(X)]$ for a complicated f.
Assume
  1. $f(x)$ is twice differentiable in x
  2. $\mu=\mathbb{E}[X]$ and $\sigma^2=\mathbb{V}[X]$ are finite

  Then by Taylor series,
$$\mathbb{E}[f(X)]\approx f(\mu)+\frac{\sigma^2}{2}\frac{d^2}{dx^2}f(x)\Big|_{x=\mu},\qquad \mathbb{V}[f(X)]\approx\sigma^2\frac{d}{dx}f(x)\Big|_{x=\mu}.$$

- **Weak law of large numbers**
  – Given n samples $\boldsymbol{x}=\{x_1,\dots,x_n\}$ of a RV X with population mean μ
  – Sample mean: $\bar{x}:=\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$
  – $\bar{x}\to\mu$ in probability when $n\to\infty$,
  I.e., for any $\epsilon>0$ we have $\displaystyle\lim_{n\to\infty}\mathbb{P}(|\bar{x}-\mu|\leq\epsilon)=1$
  In words: the more samples (bigger n), the higher chance $\bar{x}$ is the same as population mean

3.2 Covariance and correlation

- **Covariance** $\mathrm{cov}(X,Y)$ tells how much X,Y varies together

$$\mathrm{cov}(X,Y) \;:=\; \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \;=\; \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y].$$

  - The range of covariance is from $-\infty$ to $+\infty$, it depends on the scale of the variable
  - Positive cov: if $X > \mathbb{E}[X]$ then likely $Y > \mathbb{E}[Y]$
  - Negative cov: if $X > \mathbb{E}[X]$ then likely $Y < \mathbb{E}[Y]$
  - If X,Y are independent, $\mathrm{cov}(X,Y)=0$
  - Given two data vectors $\boldsymbol{x}=\{x_1,x_2,...\}$, $\boldsymbol{y}=\{y_1,y_2,...\}$, the *empirical covariance* can be written in vector form

$$\mathrm{cov}(X,Y) := \frac{1}{n}(\boldsymbol{x}-\mu_X\mathbf{1})^\top(\boldsymbol{y}-\mu_Y\mathbf{1}). \qquad \text{(empirical covariance)}$$

    where $\mathbf{1}$ is vector of all one.
    Here

    * We only have access to the data $\boldsymbol{x},\boldsymbol{y}$
    * We do not have access to the joint probability $\mathbb{P}(X=x,Y=y)$
    * Here we are assuming $\mathbb{P}(X=x_1,Y=y_1)=\mathbb{P}(X=x_2,Y=y_2)=...=\mathbb{P}(X=x_n,Y=y_n)$ and this gives the term $\dfrac{1}{n}$.
    * The assumption $\mathbb{P}(X=x_1,Y=y_1)=\mathbb{P}(X=x_2,Y=y_2)=...=\mathbb{P}(X=x_n,Y=y_n)$ is *empirical*, meaning that it may not be true, so strictly speaking we have

$$\mathrm{cov}(X,Y) \approx \mathrm{cov}_{\textsf{empirical}}(\boldsymbol{x},\boldsymbol{y}) := \frac{1}{n}(\boldsymbol{x}-\mu_X\mathbf{1})^\top(\boldsymbol{y}-\mu_Y\mathbf{1}). \qquad \text{(empirical covariance)}$$

      Most people don't care and don't distinguish between cov and $\mathrm{cov}_{\textsf{empirical}}$.
    * Furthermore, in practise we do not know the value of (μ_X,μ_Y), we take the approximation $\mu_x \approx \overline{x}$ and $\mu_y \approx \overline{y}$, and we call

$$\mathrm{cov}(X,Y) = \underbrace{\frac{1}{n}(\boldsymbol{x}-\overline{x}\mathbf{1})^\top(\boldsymbol{y}-\overline{y}\mathbf{1})}_{\text{empirical covariance with sample means}}.$$

    * Strictly speaking covariance $\neq$ empirical covariance, however many people don't care.

Example Suppose there are $n=5$ students, who spent $\{3,5,2,7,4\}$ hours to study before the exam, and got grades $\{70,80,60,90,75\}$. Find the covariance between $X=\{$the number of hours of study$\}$ and $Y=\{$grade$\}$.

Solution Let $\boldsymbol{x}=\{3,5,2,7,4\}$ and $\boldsymbol{y}=\{70,80,60,90,75\}$. The sample mean (average) of $\boldsymbol{x}$, denoted as $\overline{x}$, is $\displaystyle\frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{5}\Big(3+5+2+7+4\Big)=4.2$. The sample mean of $\boldsymbol{y}$, is $\displaystyle\overline{y}=\frac{1}{n}\sum_{i=1}^n y_i = \frac{1}{5}\Big(70+80+60+90+75\Big)=75$.

$$\begin{aligned}
\mathrm{cov}(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]=\frac{1}{n}(\boldsymbol{x}-\overline{x}\mathbf{1})^\top(\boldsymbol{y}-\overline{y}\mathbf{1}) \quad &= \quad \frac{1}{n}\begin{bmatrix} x_1-\overline{x}\\ x_2-\overline{x}\\ \vdots \end{bmatrix}^\top \begin{bmatrix} y_1-\overline{x}\\ y_2-\overline{x}\\ \vdots \end{bmatrix}\\
&= \quad \frac{1}{5}\begin{bmatrix} 3-4.2\\ 5-4.2\\ 2-4.2\\ 7-4.2\\ 4-4.2 \end{bmatrix}^\top \begin{bmatrix} 70-75\\ 80-75\\ 60-75\\ 90-75\\ 75-75 \end{bmatrix} = \frac{85}{5} = 21.25.
\end{aligned}$$

The positive covariance suggest a positive association between the number of hours studied and grade.

- It is only an association result
- It is not a causation result: it didn't say that "if you study longer, you get higher grade"

Remark Can we calculate $\mathrm{cov}(X,Y) := \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$ here? The answer is no because $\mathbb{E}[XY]$ requires the information of the joint distribution $p(x,y)=\mathbb{P}(X=x,Y=y)$, which is NOT provided here.

- **Correlation** (normalized covariance)
$$\mathrm{corr}(X,Y) := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}.$$
In practise we do not know the value of (σ_X, σ_Y), we take the approximation $\sigma_X \approx s_{\boldsymbol{x}}$ and $\sigma_Y \approx s_{\boldsymbol{y}}$, and we call
$$\mathrm{corr}(X,Y) := \frac{\mathrm{cov}(X,Y)}{s_{\boldsymbol{x}}s_{\boldsymbol{y}}}.$$
empirical correlation. Again correlation $\neq$ empirical correlation.
  - The range of covariance is from -1 to $+1$, it is independent of the scale of the variable
  - Positive corr: if $X > \mathbb{E}[X]$ then likely $Y > \mathbb{E}[Y]$
  - Negative corr: if $X > \mathbb{E}[X]$ then likely $Y < \mathbb{E}[Y]$
- **Example** For $\boldsymbol{x} = \{3,5,2,7,4\}$ and $\boldsymbol{y} = \{70,80,60,90,75\}$, we have $\mathbb{V}[X] = 3.7$ and $\mathbb{V}[Y] = 125$ The correlation is
$$\mathrm{corr}(X,Y) := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \frac{21.25}{\sqrt{3.7}\sqrt{125}} = 0.988.$$
This value is close to 1, indicating there is a strong positive association between the number of hours studied and grade. In fact, if we plot the points, we can see a clear positive trend between $\boldsymbol{x}$ and $\boldsymbol{y}$.
- **Theorem** If X,Y are independent, $\mathrm{cov}(X,Y) = \mathrm{corr}(X,Y) = 0$.
Proof $\mathrm{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \overset{\text{independent}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$ □
  - Converse not true: $\mathrm{corr}(X,Y) = 0$ does not mean X,Y are independent
  - If X,Y independent, "Variance quadratic formula" becomes $\mathbb{V}[aX \pm bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$
  Or $\mathbb{V}[X \pm Y] = \mathbb{V}[X] + \mathbb{V}[Y]$
- **Linkage between covariance and linear algebra**
Given two data vectors $\boldsymbol{x} = \{x_1, x_2, \ldots\}$, $\boldsymbol{y} = \{y_1, y_2, \ldots\}$, the empirical covariance and the empirical correlation
$$\mathrm{cov}(X,Y) = \frac{1}{n}(\boldsymbol{x} - \overline{x}\mathbf{1})^\top(\boldsymbol{y} - \overline{y}\mathbf{1}), \qquad \mathrm{corr}(X,Y) = \frac{1}{n}\frac{(\boldsymbol{x} - \overline{x}\mathbf{1})^\top(\boldsymbol{y} - \overline{y}\mathbf{1})}{s_{\boldsymbol{x}}s_{\boldsymbol{y}}},$$
If the sample mean are zero, then
$$\begin{aligned}\mathrm{corr}(X,Y) = \frac{1}{n}\frac{\boldsymbol{x}^\top\boldsymbol{y}}{s_{\boldsymbol{x}}s_{\boldsymbol{y}}} &= \frac{1}{n}\frac{\boldsymbol{x}^\top\boldsymbol{y}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n x_i^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^n y_i^2}} \\ &= \frac{n-1}{n}\frac{\boldsymbol{x}^\top\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} \\ &= \frac{n-1}{n}\frac{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2\cos\theta(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} = \frac{n-1}{n}\cos\theta(\boldsymbol{x},\boldsymbol{y})\end{aligned}$$
When n increase, the factor $\dfrac{n-1}{n} \to 1$.
Hence, what covariance means: it is the cosine angle between two data vectors $\boldsymbol{x}, \boldsymbol{y}$.

4 The normal distribution

4.1 Normal distribution

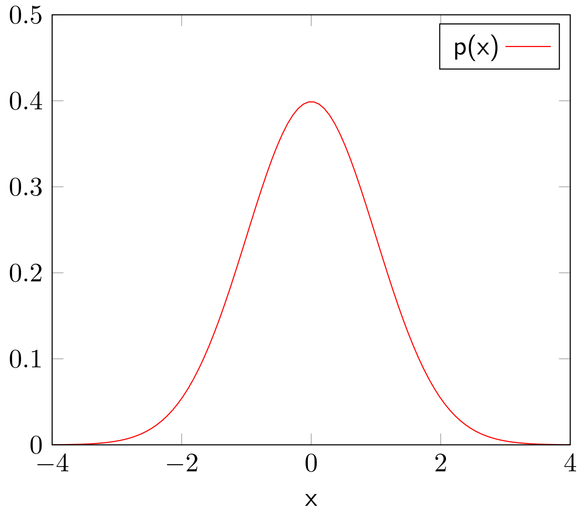
- Normal distribution is also called Gaussian distribution
- The sample space $\mathcal{X} = \mathbb{R}$ is the whole real line
- $X \sim \mathcal{N}(\mu, \sigma^2)$ means X is a RV under normal distribution with mean μ and variance σ^2

- $\mathbb{P}(X = x; \mu, \sigma^2) =: p(x|\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\dfrac{-(x-\mu)^2}{2\sigma^2}\right)$
- Probability of an interval = the area under the curve

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \int_a^b p(x)dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)dx.\end{aligned}$$

We do not compute $\int e^{-cx^2}dx$ by hand. People solve it by table lookup or by computer.

For example: $\mu = 0, \sigma = 1$



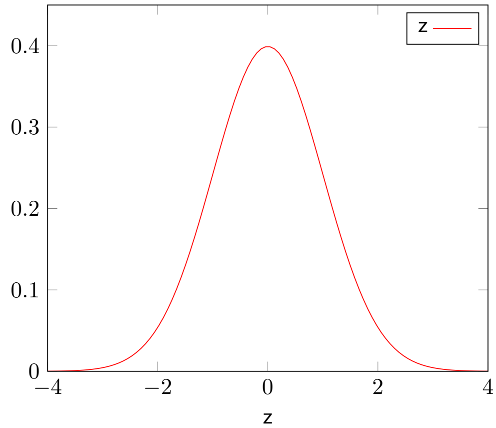
- $\mathbb{E}[X] = \mu$
- $\mathbb{V}[X] = \sigma^2$
- Normal distribution is symmetric around μ
  - mean = median = mode = μ
  - 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$ approx. 1 in 3 / thrice a week
  - 95.45% of probability falls within $(\mu - 2\sigma, \mu + 2\sigma)$ approx. 1 in 22 / every three weeks
  - 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$ approx. 1 in 370 / once a year
  - 99.99994% of probability falls within $(\mu - 5\sigma, \mu + 5\sigma)$ approx. 1 in 1744278 / once every 4776 years
- Scaling of normal variable
  If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for $\alpha > 0$ and $Y = \alpha X$, then $Y \sim \mathcal{N}(\alpha\mu, (\alpha\sigma)^2)$.
  Proof: by the variance quadratic formula, if $\sigma^2 = \mathbb{V}[X]$ and $\alpha > 0$ then $\mathbb{V}[cX] = \alpha^2\mathbb{V}[X]$
- Shifting of normal variable
  If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any c and $Y = X - c$, then $Y \sim \mathcal{N}(\mu - c, \sigma^2)$.
  Proof: by expectation is linear, $\mathbb{E}[X - c] = \mathbb{E}[X] - c$
- Example: if $X \sim \mathcal{N}\big(\mu, \sigma^2\big)$, then
  - $Y = \dfrac{X}{2}$, then $Y \sim \mathcal{N}\Big(\dfrac{1}{2}\mu,\, \dfrac{1}{4}\sigma^2\Big)$.
  - $Y = \dfrac{X}{\sigma}$, then $Y \sim \mathcal{N}\Big(\dfrac{1}{\sigma}\mu,\, 1\Big)$.
  - $Y = X - \mu$, then $Y \sim \mathcal{N}\big(0,\, \sigma^2\big)$.
  - $Y = \dfrac{X-\mu}{\sigma}$, then $Y \sim \mathcal{N}\big(0,\, 1\big)$, this process is also called standardization.
- Property of normal sum
  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. If X_1, X_2 are independent, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  The proof is out of the scope of this course.
- **Theorem** For $i = 1, 2, ..., n$, if $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, 2, ..., n$ are independent, then

$$Y = \sum c_i X_i \sim \mathcal{N}\Big(\sum c_i\mu_i, \sum c_i\sigma_i^2\Big)$$

  The proof is out of the scope of this course.
- (Chi-square) If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = X^2$ is not a normal variable but a chi-square variable. We write $X^2 \sim \chi^2(\nu)$.
  We do not talk about chi-square in this course.

4.2 Standard normal distribution

- Standard norm distribution is when $\mu = 0$ and $\sigma = 1$



  All normal distribution is a translated and scaled version of $\mathcal{N}(0,1)$

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

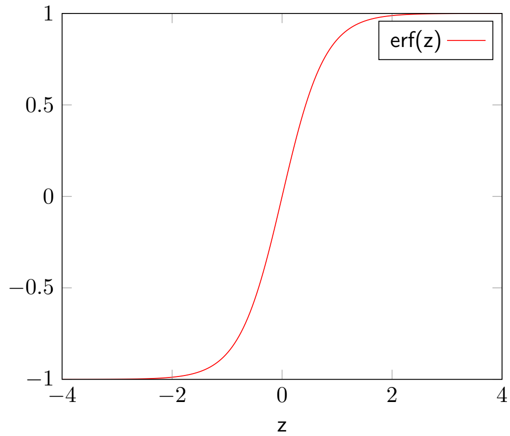
  If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

  The process $\dfrac{X-\mu}{\sigma}$ is called standardization

  If $Z \sim \mathcal{N}(0,1)$ then we call the random variable standard z score

- Calculation of z-score: recall that probability of an interval = the area under the PDF curve

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}dx.$$



  We can solve this integral by the **error function**

$$\text{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt.$$

  Note that

  - $\text{erf}(-z) = \text{erf}(z)$

  The details: we perform change of variable $t = \dfrac{x}{\sqrt{2}}$

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-(x/\sqrt{2})^2}dx &= \frac{1}{\sqrt{\pi}} \int_a^b e^{-(x/\sqrt{2})^2} \frac{dx}{\sqrt{2}} \\
&= \frac{1}{\sqrt{\pi}} \int_a^b e^{-t^2} dt \\
&= \frac{1}{\sqrt{\pi}} \int_0^b e^{-t^2} dt - \frac{1}{\sqrt{\pi}} \int_0^a e^{-t^2} dt \\
&= \frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int_0^b e^{-t^2} dt - \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt \right) = \frac{1}{2}\Big(\text{erf}(b) - \text{erf}(a)\Big)
\end{aligned}$$

  How do we calculate the error function: in the old days, people use the z-table. Nowadays, use computer!

- Using WolframAlpha `https://www.wolframalpha.com/`
  - Example:

```
integrate 1/(sqrt(2 pi)) exp(  -x^2/2 ) dx  for x = - infinity to x = 3
```

    will compute

$$\int_{-\infty}^3 \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{x^2}{2} \Big\} dx.$$

- We compute normal distribution by using standard normal distribution

- We calculate everything on $z \sim Z$ using error function
- We translate and scale back to the RV X we want to study
- **Example** If $Z \sim \mathcal{N}(0,1)$, find $\mathbb{P}(-1 \leq z \leq 1)$

$$\mathbb{P}(-1 \leq z \leq 1) \;=\; \frac{1}{2}\Big(\mathrm{erf}(1) - \mathrm{erf}(-1)\Big) \;=\; \frac{1}{2}\Big(\mathrm{erf}(1) + \mathrm{erf}(1)\Big) \;=\; \mathrm{erf}(1) = 0.84270079295$$

5 Other common distributions

5.1 Bernoulli distribution

- $\mathcal{X} = \{0,1\}$ and $x \in \mathcal{X}$ represents success or fail, any binary event
  - Coin flip (H=0, T=1)
  - Manufacturing: defects, not defects
  - Medicine: disease, no disease
  - Sport: win, lose, assume there is no draw
- $X \sim$ Ber(θ) means X is a RV under Bernoulli distribution with probability of success θ
- $\theta \in [0,1]$ represent the probability of success
  - in coin flip, a coin is fair if $\theta = 0.5$
  - in medicine, you want θ as close to 0 as possible (low chance to have disease)
- $\mathbb{P}(X = x|\theta) =: p(x|\theta) = \theta^x(1-\theta)^{1-x}$
  - It may seems wrong we have to multiply θ with $(1-\theta)$, but note that for their power x and $1-x$, only one of them is nonzero.
  - If $x = 1$ then $p(1) = \theta$
  - If $x = 0$ then $p(0) = 1 - \theta$
- $\mathbb{E}[X] = \theta$
- $\mathbb{V}[X] = \theta(1-\theta)$

Rademacher distribution If $X \sim$ Ber(0.5), then $Y = 2X - 1$ is follows Rademacher distribution, which is useful to model $Y = \{-1, +1\}$, which is very useful for modelling random walk (you either move forward $(x = +1)$ or backward $(x = -1)$, not moving $(x = 0)$ is not allowed)

5.2 Binomial distribution

- n binary RV $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ where all $X_i \sim$ Ber(θ)
- We now counts the number of success in this n binary event.

$$\text{number of success} = m = \sum_{i=1}^{n} x_i$$

- The sample space of m is then the set of integers $\{0,1,2,...,n\} =: \mathcal{M}$
- $p(m|\theta) = \binom{n}{m}\theta^m(1-\theta)^{n-m}$ is the probability that M takes a particular m success of n binary RV
- $\mathbb{E}[X] = n\theta$
- $\mathbb{V}[X] = n\theta(1-\theta)$
- Property of binomial sum
  - $M_1 \sim$ Bin(θ, n_1) and $M_2 \sim$ Bin(θ, n_2) then $M_1 + M_2 \sim$ Bin$(\theta, n_1 + n_2)$

5.3 Multinomial distribution

- Multinomial distribution generalizes the binomial distribution to independent random experiments with more than two outcomes
- $\boldsymbol{x} = (x_1, ..., x_m)$ is the random vector, the $\theta = (\theta_1, \theta_2, ..., \theta_m)$ is the success probability vector, then the multinomial probability of $\boldsymbol{x}$ achieving k_1-success, k_2-success, ..., k_m-success is then

$$\mathbb{P}\Big(\boldsymbol{x} = (k_1, k_2, ..., k_m)\Big) \;=\; \binom{n}{k_1, k_2, ..., k_m} \prod_{i=1}^{m} \theta_i^{k_i} \;=\; \frac{n!}{k_1! k_2! \dots k_m!} \theta_1^{k_1} \theta_2^{k_2} ... \theta_m^{k_m}$$

where $k_1 + k_2 + ... + k_m = n$.

5.4 Discrete uniform distribution

- $\mathcal{X} = \{a, ..., b\}$, the sample space is an integer interval from a to b
- $X \sim U(a, b)$ means X is a RV under discrete uniform distribution
- $p(X = k; a, b) = \dfrac{1}{b - a + 1}$ represent the probability X takes the value k in $\mathcal{X}$
- $\mathbb{E}[X] = \dfrac{a + b}{2}$, possibly not an integer
- $\mathbb{V}[X] = \dfrac{(b - a + 1)^2 - 1}{12}$, possibly not an integer

5.5 Poisson distribution

- $\mathbb{Z}_+ = \{0, 1, 2, ..\}$, the sample space is all nonnegative integers (from 0 to ∞)
- $X \sim$ Poi(λ) means X is a RV under discrete Poisson distribution
- $p(X = k|\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}$ represent the probability k events occurred under rate λ
- $\mathbb{E}[X] = \lambda$
- $\mathbb{V}[X] = \lambda$
- Property of Poisson sum
  - $X_1 \sim$ Poi(λ_1) and $M_2 \sim$ Poi(λ_2) then $X_1 + X_2 \sim$ Poi$(\lambda_1 + \lambda_2)$

5.6 Negative binomial distribution

- Binomial variable Bin(θ, n) then $p(m|\theta)$ refers to probability of $\boxed{\text{within } n \text{ trial, there are exactly } m \text{ success}}$
- Negative binomial refers to the probability $\boxed{\text{until the } r\text{th success}}$
- $X \sim$ NB$(r|\theta, n)$ has $p(r) = \dbinom{n + r - 1}{n} (1 - \theta)^n \theta^r$

Geometric distribution When $r = 1$, we have the geometric distribution

5.7 Special thing on continuous distribution: zero point-wise probability

- If a RV X follows a continuous distribution, then the probability $\mathbb{P}(x|\theta)$ for a particular x is always zero
- Example: $p(1) = p(0) = p(-2) = p(e) = p(\pi) = 0$
- This is because the sample space of a continuous distribution has infinitely many elements, so the chance of randomly picking a particular element is zero

- **An important consequence: strict inequality is the same as non-strict inequality**.
  That is, $\mathbb{P}(X \le a) = \mathbb{P}(X < a)$

$$
\begin{array}{rcll}
\mathbb{P}(X \le a) & = & \mathbb{P}(X < a \text{ OR } X = a) & \\
 & = & \mathbb{P}(X < a) + \mathbb{P}(X = a) & \text{inclusion-exclusion principle / sum rule / } \sigma\text{-additivity} \\
 & = & \mathbb{P}(X < a) + 0 & \mathbb{P}(X = a) \equiv 0 \text{ for any } a \\
 & = & \mathbb{P}(X < a) &
\end{array}
$$

- **A crazy fact**. For continuous random variable, $p(a) \equiv 0$ but it doesn't mean event $X = a$ is impossible.

5.8 Continuous uniform distribution

- $\mathcal{X} = [a, ..., b]$, the sample space is an interval of real number from a to b
- $X \sim U(a, b)$ means X is a RV under continuous uniform distribution
- $p(x; a, b) = \begin{cases} 0 & x < a \\ \dfrac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases}$ represent the probability X takes the value k in $\mathcal{X}$
- $\mathbb{E}[X] = \dfrac{a+b}{2}$
- $\mathbb{V}[X] = \dfrac{(b-a)^2}{12}$,

5.9 Exponential distribution

- $\mathcal{X} = [0, +\infty)$, the sample space is the positive real number
- $X \sim \exp(\lambda)$ means X is a RV under exponential distribution with rate λ
- $p(x|\lambda) = \lambda e^{-\lambda x}$
- $\mathbb{E}[X] = \dfrac{1}{\lambda}$
- $\mathbb{V}[X] = \dfrac{1}{\lambda^2}$

5.10 Other advanced distributions

- Hyper-geometric
- Gamma distribution
- Cauchy
- Beta
- Chi-squared

6 Point estimation: maximum likelihood estimator

- **The motivation**: suppose we have observed data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. Now we would like to model these using a normal distribution $p(y|\mu, \sigma^2)$, where the *population* μ, σ^2 are unknown. The process of *point estimation* is to estimate these population parameter.
- There are several approaches here
  1. Minimum Sum of Squared Errors
  2. Maximum likelihood estimator
  3. Unbiased estimator
- Notation: θ denotes the ground truth population parameter, $\hat{\theta}$ denotes the estimator

6.1 Minimum Sum of Squared Errors

- In this approach we find $\hat{\theta}$ that "close" to all data point
- Suppose we want to learn μ in $p(y|\mu, \sigma^2)$.
- The notion of "closeness" here is $\text{SSE}(\mu) := \sum_{i=1}^{n} (y_i - \mu)^2$
- We find $\hat{\mu}$ as the minimizer of $\text{SSE}(\mu)$

$$\hat{\mu} = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2$$

If we take the derivative to zero, it gives

$$\frac{d\text{SSE}(\mu)}{d\mu} = -2\sum_{i=1}^{n} y_i + 2n\mu = 0. \quad \implies \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} =: \text{sample mean}$$

6.2 Maximum likelihood

- Maximum likelihood is a popular method in parameter estimation
- In this approach we find $\hat{\theta}$ that has the highest probability given the data
- We are given $\boldsymbol{y}$, we want to find θ that maximize $p(y|\theta)$, called likelihood

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(\theta|y)$$

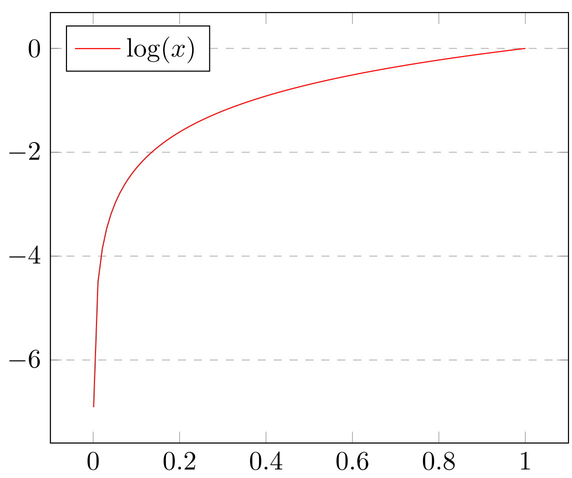
Note that

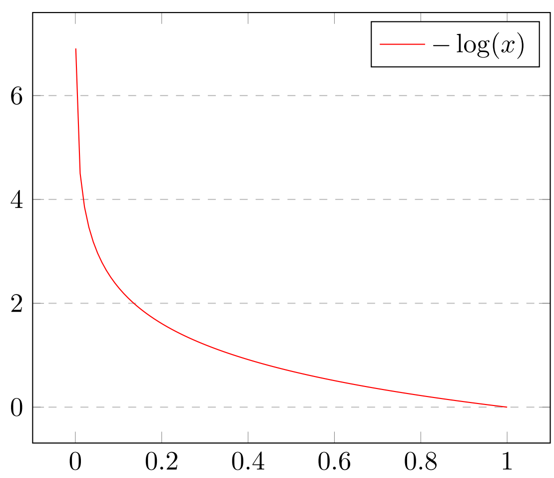
  - **here it is $p(\theta|y)$ not $p(y|\theta)$**
  - Likelihood is also a probability
  - The term "likelihood" is just a probability that "given the observed data y, how likely it is to give parameter θ"

- Due to mathematically more convenient, we sometimes work with negative log-likelihood

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \left\{ -\log p(\theta|\boldsymbol{y}) \right\}$$

Why do this: probability is a number between 0 and 1. The log "magnifies" such number from 0 to negative infinity. We multiply -1 to make the range from 0 to positive infinity.





6.2.1 Maximum Likelihood Estimation (MLE) of normal distribution

- Suppose you are given a set of observed data $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ that contains n data points $y_i,\ i \in \{1, 2, \dots, n\}$. You believe that this dataset $\boldsymbol{y}$ follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ under the following assumption.

> **IID assumption**
>
> – You assume that data points $y_1, \dots, y_n$ are i.i.d. (Independent and identically distributed) under a normal distribution $\mathcal{N}(\mu, \sigma^2)$. I.e., each data point y_i is an *realization* of a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$
>
> – In other words, each $y_1, \dots, y_n$ are random sample from a population that is normally distributed with mean μ and variance σ^2

- You don't know the population parameter μ, σ and you want to estimate μ and σ from data $\boldsymbol{y}$. There are many approaches to estimate μ, σ, now you choose to use MLE

- The MLE process starts with the likelihood function. The likelihood of y_i sampled from $\mathcal{N}(\mu, \sigma^2)$ is $p(\mu, \sigma | y_i)$. We have

$$
\begin{aligned}
p(\mu_1, \sigma_1 | y_1) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\Big(- \frac{(y_1 - \mu_1)^2}{2\sigma_1^2} \Big) \\
&\vdots \\
p(\mu_n, \sigma_n | y_n) &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\Big(- \frac{(y_n - \mu_n)^2}{2\sigma_n^2} \Big)
\end{aligned}
$$

By the IID assumption, we have $\mu_1 = \cdots = \mu_n = \mu$ and $\sigma_1 = \cdots = \sigma_n$ and therefore

$$
\begin{aligned}
p(\mu, \sigma | y_1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_1 - \mu)^2}{2\sigma^2} \Big) \\
&\vdots \\
p(\mu, \sigma | y_n) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_n - \mu)^2}{2\sigma^2} \Big)
\end{aligned}
$$

By product rule, the likelihood of observing data $\boldsymbol{y}$ from $Y \sim \mathcal{N}(\mu, \sigma^2)$ is thus

$$
\begin{aligned}
p(\mu, \sigma | \boldsymbol{y}) \;=\; p(\mu, \sigma | y_1) \cdots p(\mu, \sigma | y_n) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_1 - \mu)^2}{2\sigma^2} \Big) \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_n - \mu)^2}{2\sigma^2} \Big) \\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}}}_{n \text{ of them}} \exp\Big(- \frac{(y_1 - \mu)^2}{2\sigma^2} \Big) \cdots \exp\Big(- \frac{(y_n - \mu)^2}{2\sigma^2} \Big) \\
&= \Big(\frac{1}{2\pi\sigma^2} \Big)^{\frac{n}{2}} \exp\Bigg(- \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \Bigg), \quad \text{recall } e^{-a}e^{-b} = e^{-a-b}.
\end{aligned}
$$

- The negative log-likelihood

$$
\begin{aligned}
\mathcal{L}(\mu, \sigma | \boldsymbol{y}) \;:=\; -\log p(\mu, \sigma | \boldsymbol{y}) &= -\log \Bigg\{ \Big(\frac{1}{2\pi\sigma^2} \Big)^{\frac{n}{2}} \exp\Bigg(- \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \Bigg) \Bigg\} \\
&= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2, \quad \text{recall } \log\Big(\frac{1}{a} e^{-b}\Big) = \log a + b.
\end{aligned}
$$

- We now find the optimal μ by MLE with negative log-likelihood as $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \Big\{ -\log p(\theta | \boldsymbol{y}) \Big\}$, so

$$
\hat{\mu}_{\mathsf{MLE}} \;=\; \underset{\mu}{\operatorname{argmin}}\ \mathcal{L}(\mu, \sigma | \boldsymbol{y}) \;=\; \underset{\mu}{\operatorname{argmin}}\ \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.
$$

The minimizer of $\mathcal{L}$ with respect to μ, denoted as $\hat{\mu}_{\text{MLE}}$, is at where $\dfrac{\partial \mathcal{L}}{\partial \mu}\bigg|_{\mu=\hat{\mu}_{\text{MLE}}} = 0$, which is

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu}\bigg|_{\mu=\hat{\mu}_{\text{MLE}}} = -\frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \hat{\mu}_{\text{MLE}}) = 0 \iff \sum_{i=1}^n (y_i - \hat{\mu}_{\text{MLE}}) &= 0 \\
\iff \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\mu}_{\text{MLE}} &= 0 \\
\iff \sum_{i=1}^n y_i - n\hat{\mu}_{\text{MLE}} &= 0 \\
\iff \hat{\mu}_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^n y_i =: \overline{y} = \text{sample mean}
\end{aligned}
$$

Thus, when estimating normal distribution from a data, the maximum likelihood estimator of the population mean $\hat{\mu}_{\text{MLE}} = \dfrac{1}{n}\sum_{i=1}^n y_i$ is the sample mean.

- We now find the optimal σ^2 by MLE with negative log-likelihood as $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \Big\{ -\log p(\theta|\boldsymbol{y}) \Big\}$, so

$$
\begin{aligned}
\hat{\sigma}^2_{\text{MLE}} = \underset{\sigma^2}{\operatorname{argmin}}\ \mathcal{L}(\mu, \sigma|\boldsymbol{y}) &= \underset{\sigma^2}{\operatorname{argmin}}\ \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2 \\
&= \underset{\sigma^2}{\operatorname{argmin}}\ \frac{n}{2}\log 2\pi + \frac{n}{2}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2 \frac{1}{\sigma^2}.
\end{aligned}
$$

Note that variance is σ^2 so we are considering the symbol σ^2 instead of σ (which is standard deviation).

The minimizer of $\mathcal{L}$ with respect to σ^2, denoted as $\hat{\sigma}^2_{\text{MLE}}$, is at where $\dfrac{\partial \mathcal{L}}{\partial \sigma^2}\bigg|_{\sigma=\hat{\sigma}_{\text{MLE}}} = 0$, which is

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma^2}\bigg|_{\sigma=\hat{\sigma}_{\text{MLE}}} = \frac{n}{\hat{\sigma}^2_{\text{MLE}}} - \sum_{i=1}^n (y_i - \mu)^2 \frac{1}{\hat{\sigma}^4_{\text{MLE}}} = 0 \iff \frac{1}{\hat{\sigma}^2_{\text{MLE}}}\left(n - \sum_{i=1}^n (y_i - \mu)^2 \frac{1}{\hat{\sigma}^2_{\text{MLE}}}\right) &= 0 \\
\iff \underbrace{\frac{1}{\hat{\sigma}^2_{\text{MLE}}} = 0}_{\text{impossible}} \text{ or } n - \sum_{i=1}^n (y_i - \mu)^2 \frac{1}{\hat{\sigma}^2_{\text{MLE}}} &= 0 \\
\iff \hat{\sigma}^2_{\text{MLE}} - \frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2 &= 0 \\
\iff \hat{\sigma}^2_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2
\end{aligned}
$$

Thus, when estimating normal distribution from a data, the maximum likelihood estimator of the population variance $\hat{\sigma}^2_{\text{MLE}}(\mu) = \dfrac{1}{n}\sum_{i=1}^n (y_i - \mu)^2$ **is a function of** μ.

  - If we do not know the population μ and we estimate it by $\hat{\mu}_{\text{MLE}}$, then the MLE of σ is the sample variance

$$
\hat{\sigma}^2_{\text{MLE}}(\hat{\mu}_{\text{MLE}}) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\mu}_{\text{MLE}})^2, \text{ where } \hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n y_i
$$

  - If we know the population μ, then the MLE of σ is the sample variance

$$
\hat{\sigma}^2_{\text{MLE}}(\mu) = \frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2.
$$

- Knowing μ vs not knowing μ has a big difference.

– If μ is known: we have $\hat{\sigma}^2_{\mathsf{MLE}}(\mu)$. Now consider itself as a random variable. Consider the $\mathbb{E}\Big[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)\Big]$

$$\mathbb{E}\Big[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)\Big] \;=\; \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2\Big]. \tag{\#}$$

Now because both $\mathbb{E}$ and $\sum$ are linear operator, so we can swap their position and get

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2\Big] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(y_i-\mu)^2\Big]. \tag{\#\#}$$

Now we focus on the term $\mathbb{E}\Big[(y_i-\mu)^2\Big]$. By the IID assumption, all y_i are realization of a random variable $Y \sim \mathcal{N}(\mu,\sigma^2)$, hence the expected value $\mathbb{E}[y_i]$ is like $\mathbb{E}[Y]$ and $\mathbb{E}\big[(y_i-\mu)^2\big]$ is like $\mathbb{E}\big[(Y-\mu)^2\big]$. Recall that $\mathbb{E}\big[(Y-\mu)^2\big]$ is the definition of variance of Y, and therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(y_i-\mu)^2\Big] \;=\; \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \;=\; \frac{1}{n}\cdot n\sigma^2 \;=\; \sigma^2. \tag{\#\#\#}$$

Now combine the three Equations (#), (##) and (###) we have

$$\mathbb{E}\Big[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)\Big] \overset{(\#)}{=} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2\Big] \overset{(\#\#)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(y_i-\mu)^2\Big] \overset{(\#\#\#)}{=} \sigma^2.$$

This equation means that if we treat $\hat{\sigma}^2_{\mathsf{MLE}}(\mu)$ itself as an random variable, then the expected value of $\hat{\sigma}^2_{\mathsf{MLE}}(\mu)$ is exactly the population variance σ^2.

Recall that in Section 3 that $\mathbb{E}$ is linear: $\mathbb{E}[aX+c] = a\mathbb{E}[X]+c$, so for $\mathbb{E}[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)] = \sigma^2$

* The $\mathbb{E}$ is taken with respect to $\hat{\sigma}_{\mathsf{MLE}}$,
* σ^2 is a constant for the $\mathbb{E}$,

thus we have

$$\mathbb{E}[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)] = \sigma^2 \iff \mathbb{E}[\hat{\sigma}^2_{\mathsf{MLE}}(\mu)] - \sigma^2 = 0 \iff \mathbb{E}\big[\hat{\sigma}^2_{\mathsf{MLE}}(\mu) - \sigma^2\big] = 0.$$

The expression $\mathbb{E}\big[\hat{\theta}-\theta\big]$ is known as the *bias* of an estimator.

In other words, we call that, when estimating the unknown population variance σ^2 of a normal distribution using observed data $\boldsymbol{y}$, the maximum likelihood estimator $\hat{\sigma}^2_{\mathsf{MLE}}(\mu)$ is an *unbiased estimator*.

– If μ is unknown and we use $\hat{\mu}_{\mathsf{MLE}}$ to estimate σ^2, we will have

$$\mathbb{E}[\hat{\sigma}^2_{\mathsf{MLE}}(\hat{\mu}_{\mathsf{MLE}})] \;=\; \frac{n-1}{n}\sigma^2 \;=\; \sigma^2 - \frac{1}{n}\sigma^2.$$

* The proof is not mathematically hard but very long and is out of the scope here.
* The above expression means that the maximum likelihood estimator of variance, which is the sample variance

$$\text{sample variance } = \frac{1}{n}\sum_{i=1}^{n}\big(y_i-\overline{y}\big)^2$$

is always *under-estimating* the population variance.
* The *unbiased estimator* of population variance is actually

$$\frac{1}{n-1}\sum_{i=1}^{n}\big(y_i-\overline{y}\big)^2$$

That is, when computing the sample variance, if you count one data point less in the term n, the result is unbiased.

6.2.2 Maximum Likelihood Estimation (MLE) of Poisson distribution

- Suppose you are given a set of observed data $\boldsymbol{y} = (y_1, \dots, y_n)$ that contains n data points $y_i, i \in \{1, 2, \dots, n\}$. You believe that this dataset $\boldsymbol{y}$ follows a Poisson distribution $\mathrm{Poi}(\lambda)$

  under the following assumption.

  IID assumption

  - You assume that data points $y_1, \dots, y_n$ are i.i.d. (Independent and identically distributed) under a Poisson distribution $\mathrm{Poi}(\lambda)$. I.e., each data point y_i is an *realization* of a random variable $Y \sim \mathrm{Poi}(\lambda)$
  - I.e., each $y_1, \dots, y_n$ are random sample from a population that is Poisson distributed with rate λ

- You don't know the population parameter λ and you want to estimate λ from the data $\boldsymbol{y}$ by MLE.
- The MLE process starts with the likelihood function. The likelihood of y_i sampled from $\mathrm{Poi}(\lambda)$ is $p(\lambda|y_i)$. We have

$$\begin{aligned} p(\lambda|y_1) &= \frac{\lambda^{y_1}e^{-\lambda}}{y_1!} \\ &\vdots \\ p(\lambda|y_n) &= \frac{\lambda^{y_n}e^{-\lambda}}{y_n!} \end{aligned}$$

  Note that we have make use of the IID assumption that all y_i is sampled from the same $\mathrm{Poi}(\lambda)$ under the same rate λ.

- By product rule, the likelihood of observing data $\boldsymbol{y}$ from $Y \sim \mathrm{Poi}(\lambda)$ is thus

$$\begin{aligned} p(\mu, \sigma|\boldsymbol{y}) \;=\; p(\mu, \sigma|y_1) \cdots p(\mu, \sigma|y_n) \quad &= \frac{\lambda^{y_1}e^{-\lambda}}{y_1!} \cdots \frac{\lambda^{y_n}e^{-\lambda}}{y_n!} \\ &= \underbrace{e^{-\lambda} \cdots e^{-\lambda}}_{n \text{ of them}} \frac{\lambda^{y_1}}{y_1!} \cdots \frac{\lambda^{y_n}}{y_n!} \\ &= e^{-n\lambda} \frac{\lambda^{y_1 + \cdots + y_n}}{y_1! y_2! \dots y_n!}. \end{aligned}$$

- The negative log-likelihood

$$\begin{aligned} \mathcal{L}(\mu, \sigma|\boldsymbol{y}) \;:=\; -\log p(\mu, \sigma|\boldsymbol{y}) \quad &= \quad -\log \left\{ e^{-n\lambda} \frac{\lambda^{y_1 + \cdots + y_n}}{y_1! y_2! \dots y_n!} \right\} \\ &= \quad -\log \left\{ e^{-n\lambda} \right\} - \log \left\{ \lambda^{\sum_{i=1}^n y_i} \right\} + \log \left\{ \prod_{i=1}^n y_i! \right\} \\ &= \quad n\lambda - \left(\sum_{i=1}^n y_i \right) \log \lambda + \sum_{i=1}^n \log y_i! \end{aligned}$$

- We now find the optimal λ by MLE with negative log-likelihood as $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \Big\{ -\log p(\theta|\boldsymbol{y}) \Big\}$, so

$$\begin{aligned} \hat{\lambda}_{\mathsf{MLE}} = \underset{\lambda}{\operatorname{argmin}}\; \mathcal{L}(\lambda|\boldsymbol{y}) \quad &= \quad \underset{\lambda}{\operatorname{argmin}}\; n\lambda - \left(\sum_{i=1}^n y_i \right) \log \lambda + \sum_{i=1}^n \log y_i! \\ &= \quad \underset{\lambda}{\operatorname{argmin}}\; n\lambda - \left(\sum_{i=1}^n y_i \right) \log \lambda \end{aligned}$$

  It is in the form of $\underset{x}{\operatorname{argmin}}\; f(x) = ax - b \log x$. Taking the derivative of f with respect to x to zero gives $a - \dfrac{b}{x} = 0$, which is $x = \dfrac{b}{a}$. Hence

$$\hat{\lambda}_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^n y_i = \overline{y} =: \text{sample mean}$$

  Thus, when estimating Poisson distribution from a dataset, the maximum likelihood estimator of the population rate $\hat{\lambda}_{\mathsf{MLE}} = \dfrac{1}{n} \sum_{i=1}^n y_i$ is the sample mean.

6.2.3 Maximum Likelihood Estimation (MLE) of Bernoulli distribution

- After illustrating the MLE process for normal distribution and Poisson distribution, we now repeat the same procedure for Bernoulli distribution but with a faster pace.
- Suppose we have a dataset $\boldsymbol{y} = (y_1, \dots, y_n)$ that are iid under a Bernoulli distribution Ber(θ). For example, you are tossing a coin n times and you want to know is the coin a fair coin. The likelihood for one particular tossing result is $p(\theta|y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$, and the likelihood for the n tossing result is

$$p(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}.$$

The negative log-likelihood is then

$$\mathcal{L}(\theta|\boldsymbol{y}) = -\log \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = -\left(\sum_{i=1}^{n} y_i\right)\log\theta - \left(\sum_{i=1}^{n}(1-y_i)\right)\log(1-\theta)$$

Take the derivative with respect to θ to zero gives $-\dfrac{\sum_{i=1}^{n} y_i}{\theta} + \dfrac{\sum_{i=1}^{n}(1-y_i)}{1-\theta} = 0$, that is

$$\theta\sum_{i=1}^{n}(1-y_i) = (1-\theta)\sum_{i=1}^{n} y_i \iff \theta\Big(n - \sum_{i=1}^{n} y_i\Big) = \sum_{i=1}^{n} y_i - \theta\sum_{i=1}^{n} y_i \iff \theta = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} =: \text{sample mean}$$

That is, the maximum likelihood estimator of Bernoulli distribution is the sample mean.

- Thus, by MLE, to tell a coin is fair, you toss it n times and take the sample mean, if the result is close to 0.5 then the coin is fair.

6.2.4 Maximum Likelihood Estimation (MLE) of Binomial distribution

- Suppose we have a dataset $\boldsymbol{y} = (y_1, \dots, y_n)$ of n iid trial under a binomial distribution Bin(θ, N). The likelihood is

$$p(\theta|\boldsymbol{y}) = \prod_{i=1}^{n}\binom{N}{y_i}\theta^{y_i}(1-\theta)^{N-y_i} = \Big(\prod_{i=1}^{n}\binom{N}{y_i}\Big)\theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{nN-\sum_{i=1}^{n} y_i}$$

The negative log-likelihood is then

$$\mathcal{L}(\theta|\boldsymbol{y}) = -\sum_{i=1}^{n}\binom{N}{y_i} - \Big(\sum_{i=1}^{n} y_i\Big)\log\theta - \Big(nN - \sum_{i=1}^{n} y_i\Big)\log(1-\theta).$$

Take the derivative with respect to θ to zero gives $-\dfrac{\sum_{i=1}^{n} y_i}{\theta} + \dfrac{nN - \sum_{i=1}^{n} y_i}{1-\theta} = 0$, that is

$$\theta nN - \theta\sum_{i=1}^{n} y_i = (1-\theta)\sum_{i=1}^{n} y_i \iff \theta nN - \theta\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i - \theta\sum_{i=1}^{n} y_i \iff \theta = \frac{1}{N}\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{N}\overline{y}$$

6.2.5 Maximum Likelihood Estimation (MLE) of exponential distribution

- Suppose we have a dataset $\boldsymbol{y} = (y_1, \dots, y_n)$ of n iid trial under a exponential distribution exp(λ). The likelihood is

$$p(\lambda|\boldsymbol{y}) = \prod_{i=1}^{n}\left(\lambda e^{-\lambda y_i}\right) = \lambda^n e^{-\lambda\sum_{i=1}^{n} y_i}$$

The negative log-likelihood is then

$$\mathcal{L}(\lambda|\boldsymbol{y}) = -n\log\lambda + \lambda\sum_{i=1}^{n} y_i$$

Take the derivative with respect to θ to zero gives $-\dfrac{n}{\lambda} + \sum_{i=1}^{n} y_i = 0$, that is $\lambda = \overline{y}^{\,-1}$.

7 Finite-sample statistics

7.1 The what and why of finite-sample statistics

- Finite-sample statistics refers to the behaviour of the estimator under repeated sampling.
- Finite-sample statistics is also called sampling statistics (a confusing name!)
- Why finite-sample statistics: it has three applications
  - Comparing the quality of estimators: bias and variance
  - Quantifying the accuracy of an estimator: confidence interval
  - Determine how unlikely a statistics is: hypothesis testing

  In this section we only focus on what is finite-sample statistics

7.2 Finite-sample statistics of sample mean

- Suppose we draw 5 samples from a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ three times.

The draw	Data observed	sample mean
First draw	$\boldsymbol{y}_1 = (1.62, 1.65, 1.62, 1.47, 1.62)$	$\overline{y}_1 = 1.596$
Second draw	$\boldsymbol{y}_2 = (1.72, 1.51, 1.41, 1.50, 1.68)$	$\overline{y}_2 = 1.564$
Third draw	$\boldsymbol{y}_3 = (1.68, 1.69, 1.63, 1.66, 1.60)$	$\overline{y}_3 = 1.652$

- Finite-sample statistics is asking the following question:

  what is the distribution of these $\overline{y}_1, \overline{y}_2, \overline{y}_3$?

  That is, we are now treating $\overline{y}_1, \overline{y}_2, \overline{y}_3$ as a realization of a random variable $\overline{y}$, and ask what is the statistics of $\overline{y}$.
- Now suppose we draw n sample $y_1, y_2, \ldots, y_n$ from the population $Y \sim \mathcal{N}(\mu, \sigma^2)$.
- The sample mean $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n y_i$.
- Now under iid assumption, all y_i comes from the same random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$, so if we treat the sample mean $\overline{y}$ itself as a random variable, the distribution of the sample mean $\overline{y}$ can be obtained by the property of Gaussian sum:

$$\overline{y} \coloneqq \frac{1}{n}\sum_{i=1}^n y_i \quad = \quad \frac{y_1}{n} + \frac{y_2}{n} + \cdots + \frac{y_n}{n}$$

- We recall two facts
  1. property of Gaussian sum: if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  2. Scaling of normal random variable: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\dfrac{X}{n} \sim \mathcal{N}\Big(\dfrac{\mu}{n}, \Big(\dfrac{\sigma}{n}\Big)^2\Big)$.
- Then, if we treat $\overline{y}$ as a random variable, it will be

$$\begin{aligned} \overline{y} &= \frac{y_1}{n} + \frac{y_2}{n} + \cdots + \frac{y_n}{n} \\ &\sim \mathcal{N}\Bigg(\underbrace{\frac{\mu}{n} + \cdots + \frac{\mu}{n}}_{n \text{ of them}}, \underbrace{\Big(\frac{\sigma}{n}\Big)^2 + \cdots + \Big(\frac{\sigma}{n}\Big)^2}_{n \text{ of them}} \Bigg) = \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big). \end{aligned} \qquad \text{(Distribution of sampling mean)}$$

  That means, the sample mean itself follows a normal distribution, with a mean equal to the (unknown) population mean, and a variance equal to the (unknown) population variance divided by the sample size. This means the more we take our samples, the lower the variance is $\overline{y}$.

  In other words,

  The more sample set $\boldsymbol{y}_i$ we use, the more "accurate" $\overline{y}$ in estimating μ,

  where the term "accurate" refers to $\begin{cases} \mathbb{E}[\overline{y}] \to \mu \\ \mathbb{V}[\overline{y}] \to 0 \end{cases}$

7.3 Finite-sample statistics of sample variance and chi-squared distribution

- This part is an advanced topic and not our focus / not in exam.
- Consider the unbiased estimator of population variance as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

  where $\overline{y}$ is the sample mean.
- Then by Cochrans's theorem,

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{y_i - \overline{y}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

  where χ^2_{n-1} denotes the chi-squared distribution with degree of freedom $n-1$
- The message here is that, unlike sample mean having a normal distribution, the sample variance s^2 is chi-squared distributed.
- We will skip Cochrans's theorem, chi-squared distribution and analysis of variance here.

8 Comparing estimator

Why we need to compare estimators

- Consider the following case
  - Suppose we are given dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$ and we believe the data are drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown.
  - From last section, we know that
    * the maximum likelihood estimator for μ is the sample mean $\displaystyle\frac{1}{n}\sum_{i=1}^{n} y_i$
    * the maximum likelihood estimator for σ^2 is the sample variance $\displaystyle\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$
  - We also know that the unbiased estimator for σ^2 is $\displaystyle\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

  This means we now have two estimators. How do we compare estimator: we use sampling statistics / finite-sample statistics.

How we compare estimators

- Suppose we estimate a parameter θ using two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$
- There are two things we can compare "how good are $\hat{\theta}_1$ and $\hat{\theta}_2$ on estimating θ"
  1. **Bias**: how large is the systematic error
  2. **Variance**: how large is the random error

8.1 Bias of an estimator

- Suppose we estimator a parameter θ using an estimator $\hat{\theta}$
- The bias of $\hat{\theta}$ is

$$b_\theta(\hat{\theta}) := \mathbb{E}[\hat{\theta}] - \theta. \tag{Bias}$$

- **Definition (Unbiased estimator)** If $b_\theta(\hat{\theta}) = 0$ we call $\hat{\theta}$ an unbiased estimator of θ
- What bias means: it tells that, on average, how $\hat{\theta}$ over-estimate / under-estimate θ

8.1.1 Example: bias of sample mean on the population mean of normal distribution

- Consider that:
  - given dataset $\boldsymbol{y} = (y_1, \dots, y_n)$, we believe the data are drawn from $\mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown.
  - the maximum likelihood estimator for μ is the sample mean $\overline{y} = \dfrac{1}{n}\sum_{i=1}^n y_i$

  So what is the bias of $\overline{y}$ on estimating μ?
- From Equation (Distribution of sampling mean), we see that $\overline{y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, therefore

$$b_\mu(\overline{y}) \coloneqq \mathbb{E}[\overline{y}] - \mu \overset{\mathbb{E}[\overline{y}]}{=} \mu - \mu = 0$$

  Hence the maximum likelihood estimator for μ / sampling mean, is an unbiased estimator of the population mean μ.

8.1.2 Example: bias of maximum likelihood estimator of population variance of normal distribution

- Consider that:
  - given dataset $\boldsymbol{y} = (y_1, \dots, y_n)$, we believe the data are drawn from $\mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown.
  - the maximum likelihood estimator for σ^2 is $\sigma^2_{\textsf{MLE}} = \dfrac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$ where $\overline{y} = \dfrac{1}{n}\sum_{i=1}^n y_i$ is the sample mean

  So what is the bias of $\sigma^2_{\textsf{MLE}}$ on estimating σ^2?
- The bias of $\sigma^2_{\textsf{MLE}}$ on estimating σ^2 is

$$b_{\sigma^2}(\sigma^2_{\textsf{MLE}}) = \mathbb{E}[\sigma^2_{\textsf{MLE}}] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

  In other words, we have $\mathbb{E}[\sigma^2_{\textsf{MLE}}] = \dfrac{n-1}{n}\sigma^2$, we haven't prove this one and we will not prove this one in the course.
- Instead we usually use the **unbiased estimator**

$$\hat{\sigma}^2_{\textsf{unbiased}} = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2.$$

  That is, instead of $\dfrac{1}{n}$ in $\sigma^2_{\textsf{MLE}}$, here we use $\dfrac{1}{n-1}$.

8.2 Variance of an estimator

- Suppose we estimator a parameter θ using an estimator $\hat{\theta}$
- The variance of $\hat{\theta}$ is

$$\mathbb{V}[\hat{\theta}] = \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\Big]. \tag{Var}$$

  I.e., we are treating $\hat{\theta}$ as a random variable and find the variance of $\hat{\theta}$.
- What variance means: it tells that, on average, how $\hat{\theta}$ varies around θ if we re-sampled from the population.

8.2.1 Example: variance of sample mean on the population mean of normal distribution

- Consider that:
  - given dataset $\boldsymbol{y} = (y_1, \dots, y_n)$, we believe the data are drawn from $\mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown.
  - the maximum likelihood estimator for μ is the sample mean $\overline{y} = \dfrac{1}{n}\sum_{i=1}^n y_i$

  So what is the variance of $\overline{y}$ on estimating μ?
- From Equation (Distribution of sampling mean), we see that $\overline{y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, therefore

$$\mathbb{V}[\overline{y}] = \frac{\sigma^2}{n}$$

  Hence the maximum likelihood estimator for μ / sampling mean, has a non-zero variance, and such variance decreases as n increases.

8.3 Consistency

- Suppose we estimate a parameter θ using estimator $\hat{\theta}$
- $\hat{\theta}$ is called consistent if $b_\theta(\hat{\theta}) \to 0$ and $\mathbb{V}[\hat{\theta}] \to 0$ as n increases
- The maximum likelihood estimator / sampling mean $\overline{y}$ is a consistent estimator (from the discussion above)

8.4 Mean Squared Errors

- $\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}\Big[(\hat{\theta}-\theta)^2\Big]$
  - Compared to variance of estimator $\mathbb{V}[\hat{\theta}] = \mathbb{E}\Big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\Big]$, the MSE is variance with $IE[\hat{\theta}]$ replaced by θ
- What is MSE: it tells, on average, how the estimator is awayfrom the population parameter θ
- MSE has a nice property known as bias-varaince decomposition:
  Theorem $\text{MSE}_\theta(\hat{\theta})$ can be expressed as

$$\text{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathbb{V}[\hat{\theta}] \qquad \text{(MLE bias-varaince decomposition)}$$

Proof We prove right hand side of (MLE bias-varaince decomposition) gives left hand side of (MLE bias-varaince decomposition)

$$\begin{aligned}
b_\theta^2(\hat{\theta}) + \mathbb{V}[\hat{\theta}] &= \Big(\mathbb{E}[\hat{\theta}]-\theta\Big)^2 + \mathbb{E}\Big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\Big] \\
&= \Big(\mathbb{E}[\hat{\theta}]\Big)^2 - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2 + \mathbb{E}\Big[\hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\hat{\theta}] + (\mathbb{E}[\hat{\theta}])^2\Big] \\
&= \Big(\mathbb{E}[\hat{\theta}]\Big)^2 - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2 + \mathbb{E}\Big[\hat{\theta}^2\Big] - 2\mathbb{E}\Big[\hat{\theta}\mathbb{E}[\hat{\theta}]\Big] + \mathbb{E}\Big[(\mathbb{E}[\hat{\theta}])^2\Big] \\
&= \Big(\mathbb{E}[\hat{\theta}]\Big)^2 - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2 + \mathbb{E}\Big[\hat{\theta}^2\Big] - (\mathbb{E}[\hat{\theta}])^2 \\
&= -2\mathbb{E}[\hat{\theta}]\theta + \theta^2 + \mathbb{E}\Big[\hat{\theta}^2\Big] \\
&= \mathbb{E}\Big[(\hat{\theta}-\theta)^2\Big] = \text{MSE}_\theta(\hat{\theta})
\end{aligned}$$

Remark that we have make use of the fact that $\mathbb{E}[aX] = a\mathbb{E}[X]$ and $\mathbb{E}[a] = a$

  - If the estimator is unbiased, then MSE reduecs to the variance

9 Central Limit Theorem

9.1 What is the Central Limit Theorem

- "At the limit, all random variables are normally distributed"
  This is the reason why
  - normal distribution is the most important distribution in statistics.
  - many phenomena seem to be normally distributed
- **Central Limit Theorem** Let $Y_1, Y_2, ..., Y_n$ be i.i.d. RV with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$, then

$$S = \sum_{i=1}^n Y_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \text{ as } n \to \infty$$

where $\xrightarrow{d}$ means converges in distribution.

- **De Moivre-Laplace Limit Theorem**
  - A special case of central limit theorem
  - In short, it said "we can approximate binomial distribution using normal distribution"
  - $X \sim \text{Bin}(\theta, n)$, then the standardization $\dfrac{X-n\theta}{\sqrt{n\theta(1-\theta)}}$ has the following property

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz$$

9.2 Application of CLT: approximating other distributions using normal distribution

9.2.1 Approximating binomial distribution using normal distribution

9.2.2 Approximating Poisson distribution using normal distribution

10 Interval estimation: confidence interval

- **The motivation**: suppose we have observed data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. Now we would like to model these using a distribution with the population parameter θ.
- In *point estimation*, we estimate the population parameter θ using an estimator $\hat{\theta}$.
- In *interval estimation*, we quantify how uncertain about the population parameter θ.
- In point estimation (e.g., maximum likelihood estimator), we get a single value from a given dataset $\boldsymbol{y}$. That is, we get a single value $\hat{\theta}$.
- In interval estimation (e.g., confidence interval), we get a range of value from a given dataset $\boldsymbol{y}$

$$T(\boldsymbol{y}) = (\theta^-, \theta^+) \subset \mathbb{R}$$

  which says the population parameter θ is somewhere between θ^- and θ^+.
  - θ^-: the lower bound of the interval
  - θ^+: the upper bound of the interval
  - $\theta^- \leq \theta^+$ and possibly $\theta^- = \theta^+$
  - Narrow interval: low uncertainty
    * zero interval: $\theta^- = \theta^+$, in this case we have no uncertainty
  - Wide interval: high uncertainty
- **Daily life example** You want to guess how tall (in cm) your friend is
  - Say your friend true height is 173 cm.
  - Point estimation: "it is estimated as 170 cm"
  - Interval estimation: "it is somewhere between 160 and 180 cm"
- How to obtain an interval: *confidence interval* from the frequentist statistics.

10.1 Frequentist 95% confidence interval

- $T_\alpha(\boldsymbol{y})$ is a $100(1-\alpha)\%$ confidence interval for $\alpha \in (0,1)$ if

$$\begin{aligned} & \mathbb{P}(\theta \in T_\alpha(\boldsymbol{y})) &= 1-\alpha \\ \iff & \mathbb{P}(\theta^- \leq \theta \leq \theta^+) &= 1-\alpha \end{aligned}$$

  The probability is with respect to the population distribution over all the possible data samples.
- $\alpha \in [0,1]$ is a number telling the degree of uncertainty

α	$1-\alpha$	$100(1-\alpha)\%$
0.01	0.99	99%
0.025	0.975	97.5%
0.05	0.95	95%
0.1	0.9	90%

- If $\alpha = 0.05$ we call $T_{0.05}$ the 95% confidence interval
  - It means: the probability of $\theta \in T_{0.05}(\boldsymbol{y})$ is 95%.
  - We say "we are 95% confident that θ is somewhere in $T_{0.05}(\boldsymbol{y})$
- Confidence interval **is confusing**
  - Confidence interval means **before** we draw the sample $\boldsymbol{y}$, we know that there is a 95% chance we will draw a dataset that gives an interval that covers the true θ
  - Confidence interval **does not** means **after** we draw the sample $\boldsymbol{y}$, the true θ has a 95% chance with the interval we obtained
  - The population parameter θ is not random, it is fixed

10.2 Confidence interval for normal mean, known variance

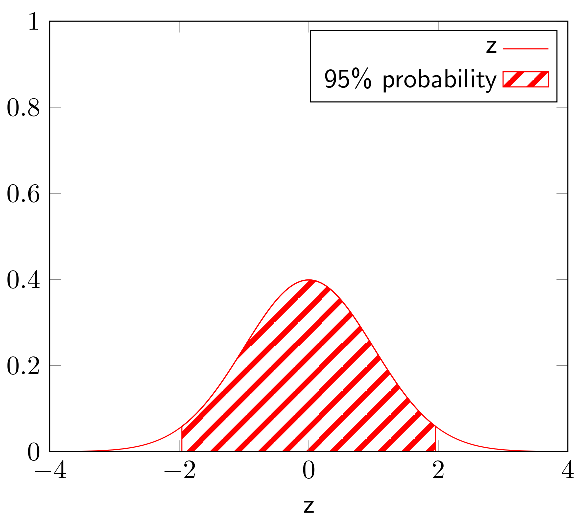
- We are given observed dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$
- We assume that data points $y_1, \ldots, y_n$ are i.i.d. (Independent and identically distributed) under a normal distribution $\mathcal{N}(\mu, \sigma^2)$. I.e., each data point y_i is an *realization* of a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$
- We assume σ^2 is known but μ is unknown
- We estimate μ by maximum likelihood estimator / sample mean $\overline{y} = \sum_{i=1}^{n} y_i$
- From the finite-sample statistics of $\overline{y}$ is that $\overline{y} \sim \mathcal{N}\Big(\mu, \dfrac{\sigma^2}{n}\Big)$.
  That is, $\overline{y}$ is a realization of a random variable $\overline{Y}$ (here the notation $\overline{Y}$ denotes a random variable, not the sample mean of Y) where $\overline{Y}$ follows a normal distribution with mean μ and variance $\dfrac{\sigma^2}{n}$
- We recall that every normal distribution is a translated and scaled version of $\mathcal{N}(0,1)$, hence

$$\frac{\overline{y}-\mu}{\sqrt{\sigma^2/n}} = \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

  $\sigma/\sqrt{n}$ has a special name called *standard error*
- Now we consider 95% conference interval
  - $\alpha = 0.05$
  - $\mathbb{P}(\theta^- \leq \theta \leq \theta^+) = 0.95 = 95\%$
  - Put θ here as the $\dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}}$, gives

$$\mathbb{P}\Big(\theta^- \leq \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq \theta^+\Big) = 0.95$$

  - Now we get the values θ^-, θ^+. Since $\dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, that means the above probability is to find the interval where the area under the curve is 0.95



    We have θ^-, θ^+ as $-1.96, +1.96$.
  - How we find the value of θ^-, θ^+: we solve $\mathbb{P}\Big(\theta^- \leq \dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq \theta^+\Big) = 0.95$ and use the fact that $z = \dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}}$ has the density function $\dfrac{1}{\sqrt{2\pi}} \exp\Big\{-\dfrac{z^2}{2}\Big\}$. Then we solve a difficult integral to find the unknonw.

$$
\begin{array}{rrcll}
 & \mathbb{P}\Big(\theta^- \leq \dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq \theta^+\Big) & = & 0.95 & \\
\iff & \displaystyle\int_{\theta^-}^{\theta^+} p(z|0,1)dz & = & 0.95 & \text{probability = area under the curve of PDF} \\
\iff & 2\displaystyle\int_{0}^{\theta^+} p(z|0,1)dz & = & 0.95 & \text{normal distribution is symmetric} \\
\iff & 2\displaystyle\int_{0}^{\theta^+} \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{z^2}{2}\Big\}dz & = & 0.95 & \text{the PDF of standard distribution} \\
\iff & \sqrt{\dfrac{2}{\pi}}\displaystyle\int_{0}^{\theta^+} \exp\Big\{-\frac{z^2}{2}\Big\}dz & = & 0.95 & \\
\iff & \mathrm{erf}\Big(\dfrac{z}{\sqrt{2}}\Big)\Big|_{z=0}^{z=\theta^+} & = & 0.95 & \text{erf is known as the error function} \\
\iff & \mathrm{erf}\Big(\dfrac{\theta^+}{\sqrt{2}}\Big) & = & 0.95 & \text{note that } \mathrm{erf}(0)=0 \\
\iff & \theta^+ & \approx & 1.95996398454005 &
\end{array}
$$

In the derivation we make use of a non-trivial fact

$$\frac{d}{dx}\mathrm{erf}\Big(\frac{x}{\sqrt{x}}\Big) = \sqrt{\frac{2}{\pi}}\exp\Big\{-\frac{x^2}{2}\Big\}$$

- In summary,

$$\mathbb{P}\Big(-1.96 \leq \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq 1.96\Big) = 0.95$$

- Recall our goal here is to obtain an interval of μ, hence we perform the following

$$
\begin{array}{rl}
 & \mathbb{P}\Big(-1.96 \leq \dfrac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq 1.96\Big) = 0.95 \\
\iff & \mathbb{P}\Big(-1.96\dfrac{\sigma}{\sqrt{n}} \leq \overline{y}-\mu \leq 1.96\dfrac{\sigma}{\sqrt{n}}\Big) = 0.95 \\
\iff & \mathbb{P}\Big(-1.96\dfrac{\sigma}{\sqrt{n}} \leq \mu-\overline{y} \leq 1.96\dfrac{\sigma}{\sqrt{n}}\Big) = 0.95 \\
\iff & \mathbb{P}\Big(\overline{y}-1.96\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{y}+1.96\dfrac{\sigma}{\sqrt{n}}\Big) = 0.95
\end{array}
$$

- Therefore, the 95% confidence interval for μ

$$T_{0.05} = \Big[\overline{y}-1.96\frac{\sigma}{\sqrt{n}},\ \overline{y}+1.96\frac{\sigma}{\sqrt{n}}\Big].$$

In other words, for 95% of the possible samples, the true population mean will be within $1.96\dfrac{\sigma}{\sqrt{n}}$ of the sample mean $\overline{y}$.

- $100(1-\alpha)\%$ confidence interval for general α
In general, you solve

$$\mathbb{P}\Big(\theta^- \leq \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq \theta^+\Big) = 1-\alpha$$

which ultimately reduce to solving

$$\mathrm{erf}\Big(\frac{z_{\alpha/2}}{\sqrt{2}}\Big) = 1-\alpha$$

with the confidence interval

$$T_\alpha = \Big[\overline{y}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{y}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big].$$

where $z_{\alpha/2}$ is the $100(1-\frac{\alpha}{2})$ percentile of the unit normal

α	$1-\frac{\alpha}{2}$	$z_{\alpha/2}$
0.01	0.995	2.576
0.025	0.9875	2.251
0.05	0.975	1.959
0.1	0.95	1.644

Example (Computing a confidence interval) You are a farmer. You grow watermelon. This season you have 8 watermelon with the following mass (in kg)

$$\boldsymbol{y} = (27.2, 7.6, 10.6, 16, 7.3, 11.8, 5.2, 17.4)$$

A watermelon biologist told you the population variance of the mass of the watermelon is 12. Find the 95% confidence interval of the population mean of the mass of the watermelon.

Solution Here we have $n = 8$ (number of data points)

The sample mean $\overline{y} = \dfrac{27.2 + 7.6 + 10.6 + 16 + 7.3 + 11.8 + 5.2 + 17.4}{8} = 12.89.$

The variance $\sigma^2 = 12$ (given) and thus $\sigma = \sqrt{12}$

For $\alpha = 0.05$, the value $z_{\alpha/2}$ such that $\mathrm{erf}\Big(\dfrac{z_{\alpha/2}}{\sqrt{2}}\Big) = 1 - \alpha$ is 1.959 (or 1.96).

Therefore, the 95% confidence interval for μ

$$T_{0.05} = \Big[\overline{y} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\Big] = \left[12.89 - 1.96 \cdot \frac{\sqrt{12}}{\sqrt{8}},\ 12.89 + 1.96 \cdot \frac{\sqrt{12}}{\sqrt{8}}\right] = [10.48, 15.29]$$

That is, the estimated mean of the mass of watermelon from your farm is 12.89 kg/melon. We are 95% confident that the population mean mass for the watermelon in your farm is between 10.48kg and 15.29kg.

Two things to note

- Confidence interval is for estimating the population mean, it is possible that the data points are not within the interval. For examples the 27.2 watermelon is not within this interval.
- Notice that the variance (=12) is large here.

Example (What does "95% confidence" mean) Suppose you obtain 5 sets of dataset draw from a random variable $\mathcal{N}(\mu = 1.65, \sigma^2 = 0.1)$. Here $\sigma^2 = 0.1$ is known and $\mu = 1.65$ is unknown and you construct 95% CI from each dataset. In each dataset, there are 4 data points, i.e., each $\boldsymbol{y}$ contains four points y_1, y_2, y_3, y_4.

Dataset	The values	Sample mean	95% CI from $\boldsymbol{y}$	$1.65 \in$ CI?
$\boldsymbol{y}_1$	$1.62, 1.80, 1.39, 1.16$	$\overline{y}_1 = 1.493$	$[1.182, 1.802]$	yes
$\boldsymbol{y}_2$	$1.64, 1.71, 2.02, 1.64$	$\overline{y}_2 = 1.753$	$[1.442, 2.062]$	yes
$\boldsymbol{y}_3$	$1.70, 1.10, 1.53, 0.90$	$\overline{y}_3 = 1.308$	$[0.997, 1.617]$	no
$\boldsymbol{y}_4$	$1.52, 1.14, 1.46, 1.45$	$\overline{y}_4 = 1.393$	$[1.082, 1.702]$	yes
$\boldsymbol{y}_5$	$1.55, 1.89, 1.63, 2.07$	$\overline{y}_5 = 1.785$	$[1.475, 2.209]$	yes
$\vdots$				

The "95%" means 5% of the time when you draw data from a population, $\mu \notin$CI

In this example, all $\overline{y}_i \neq \mu$, and the mean of the sample mean

$$\text{Avg}(\overline{y}) = \frac{\overline{y_1} + \overline{y_2} + \overline{y_3} + \overline{y_4} + \overline{y_5}}{5} = 1.564$$

is still not quite 1.65. Also,

$$\text{Var}\Big(\overline{y}\Big) = \frac{\displaystyle\sum_{i=1}^{5}\Big(\overline{y_i} - \text{Avg}(\overline{y})\Big)^2}{5} = 0.0368$$

Well, by the fact that sample mean is a random variable following $\mathcal{N}\Big(\mu, \dfrac{\sigma^2}{n}\Big)$, it tells that in this case, 5 datasets is just not enough, and if we take more datasets, then eventually Avg$(\overline{y})$ will approach $\mu = 1.65$ and Var$(\overline{y})$ will approach 0.

10.3 Confidence interval for normal mean, unknown variance

- We are given observed dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$
- We assume that data points $y_1, \ldots, y_n$ are i.i.d. (Independent and identically distributed) under a normal distribution $\mathcal{N}(\mu, \sigma^2)$. I.e., each data point y_i is an *realization* of a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$
- Unlike 10.2, now we assume both μ, σ^2 are unknown
- Since both μ, σ^2 are unknown, we need to propose two estimators to estimate them

- Like 10.2, we estimate μ by maximum likelihood estimator / sample mean $\overline{y} = \sum_{i=1}^{n} y_i$
- For estimating σ^2, things are getting complicated
  - We cannot use the maximum likelihood estimator for $\sigma^2_{\text{MLE}} = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$ due to its bias $-\dfrac{\sigma^2}{n}$
  - We use the unbiased estimator $\hat{\sigma}^2_{\text{unbiased}} = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.
  - Now using the unbiased estimator $\hat{\sigma}^2_{\text{unbiased}}$, you may think that the interval is now
  $$T_\alpha = \Big[\overline{y} - z_{\alpha/2}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}},\ \ \overline{y} + z_{\alpha/2}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}}\Big].$$
  However this does not work due to the issue that we are now estimating the variance.
  Technically, the random variable $\dfrac{\overline{y} - \mu}{\hat{\sigma}_{\text{unbiased}}/\sqrt{n}}$ is no longer normally distributed. In other words, $z_{\alpha/2}$, which comes from $\mathcal{N}(0,1)$, cannot be used.
  - Instead, the random variable $\dfrac{\overline{y} - \mu}{\hat{\sigma}_{\text{unbiased}}/\sqrt{n}}$ follows **Student-t distribution** with $n-1$ degree-of-freedom
  - Student-t distribution has a very complicated density function involving Gamma function, so we are not going to talk about it here.
- The Student-t distribution
  - "looks similar" to normal distribution
  - is also symmetric and self-similar
- The confidence interval using Student-t distribution is now
$$T_\alpha = \Big[\overline{y} - t_{\alpha/2,n-1}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}},\ \ \overline{y} + t_{\alpha/2,n-1}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}}\Big].$$
which achieves $100(1-\alpha)\%$ converge if the population is a normal random varaible.
- The value $t_{\alpha/2,n-1}$ is the $100(1-\frac{\alpha}{2})$th percentile of the standard Student-t distribution with $n-1$ degree-of-freedom. Unlike the error function, it is even more complicated to compute the $t_{\alpha/2,n-1}$
- Usually $t_{\alpha/2,n-1}$ is obtained by checking table or software (in R you run `qt( p= 1 - a/2, df = n-1)` to get it.
  - For $n = 3$, $\alpha = 0.05$, then $t_{0.025,2} = 4.3$
  - For $n = 6$, $\alpha = 0.05$, then $t_{0.025,5} = 2.57$
  - For $n = 11$, $\alpha = 0.05$, then $t_{0.025,10} = 2.22$

Example (Same watermelon example) Recall the watermelon example with
$$\boldsymbol{y} = (27.2, 7.6, 10.6, 16, 7.3, 11.8, 5.2, 17.4)$$
Find the 95% confidence interval of the population mean of the mass of the watermelon. This time we do not have the population variance.

Solution Here we have $n = 8$ (number of data points)
The sample mean $\overline{y} = \dfrac{27.2 + 7.6 + 10.6 + 16 + 7.3 + 11.8 + 5.2 + 17.4}{8} = 12.89$.
The unbiased estimator of variance $\hat{\sigma}^2_{\text{unbiased}} = \dfrac{\sum_{i=1}^{8}(y_i - 12.89)^2}{8-1} = 51.36$
For $\alpha = 0.05$, the value $t_{\alpha/2,n-1}$ is $t_{0.025,7} = 2.36$.
Therefore, the 95% confidence interval for μ
$$T_{0.05} = \Big[\overline{y} - t_{\alpha/2,n-1}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}},\ \ \overline{y} + t_{\alpha/2,n-1}\frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{n}}\Big] = \left[12.89 - 2.36\cdot\frac{\sqrt{51.36}}{\sqrt{8}},\ 12.89 + 2.36\cdot\frac{\sqrt{51.36}}{\sqrt{8}}\right] = [6.91, 18.86]$$
That is, the estimated mean of the mass of watermelon from your farm is 12.89 kg/melon, the sample variance is 51.36 (sample standard deviation is 7.166). We are 95% confident that the population mean mass for the watermelon in your farm is between 6.91kg and 18.86kg. Compared with the case with known variance, we can see now the interval is wider, because we have less information for the information (knowing σ^2 tells some information about the population and hence can be used to reduce the interval).

10.4 Confidence interval for difference of normal means

- We are given two set of data $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$, where $\boldsymbol{y}_A$ is a realization of a RV Y_A and $\boldsymbol{y}_B$ is a realization of a RV Y_B .
- We believe $Y_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, $Y_B \sim \mathcal{N}(\mu_B, \sigma_B^2)$
- We assume μ_A, μ_B are unknown and σ_A^2, σ_B^2 are known
- We assume $\boldsymbol{y}_A$ has n_A data size and $\boldsymbol{y}_B$ has n_B data size
- We now want to know is there **difference** between the two samples.
- We build a confidence interval for population mean difference $\mu_A - \mu_B$

How we build the confidence interval, the analysis

- We estimate μ_A by $\hat{\mu}_A$ and estimate μ_B by $\hat{\mu}_B$
- We can use maximum likelihood estimator, i.e., we have sample mean $\hat{\mu}_A = \overline{\boldsymbol{y}}_A$, $\hat{\mu}_B = \overline{\boldsymbol{y}}_B$
- By finite-sample statistics of mean, we have

$$\hat{\mu}_A \sim \mathcal{N}(\mu_A, \frac{\sigma_A^2}{n_A}), \qquad \hat{\mu}_B \sim \mathcal{N}(\mu_B, \frac{\sigma_B^2}{n_B})$$

- Assuming the samples are independent

$$\mathbb{V}[\hat{\mu}_A - \hat{\mu}_B] = \mathbb{V}[\hat{\mu}_A] + \mathbb{V}[\hat{\mu}_B]$$

(Recall $\mathbb{V}[aX + bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$)

- The difference $\hat{\mu}_A - \hat{\mu}_B$ staisfies

$$\hat{\mu}_A - \hat{\mu}_B \sim \mathcal{N}\Big(\mu_A - \mu_B,\ \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\Big)$$

- By $X \sim \mathcal{N}(\mu, \sigma^2) \iff Z = \dfrac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ we have

$$\hat{\mu}_A - \hat{\mu}_B \sim \mathcal{N}\Big(\mu_A - \mu_B,\ \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\Big) \quad \iff \quad \frac{(\hat{\mu}_A - \hat{\mu}_B) - (\mu_A - \mu_B)}{\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}} \sim \mathcal{N}(0, 1)$$

- Therefore we have the following $100(1 - \alpha)\%$ confidence interval

$$T_\alpha = \left[\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\ \ , \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\right]$$

Steps in computing CI for difference of normal means

1. Compute $\overline{y}_A$ from $\boldsymbol{y}_A$ and compute $\overline{y}_B$ from $\boldsymbol{y}_B$
2. Compute $z_{\alpha/2}$ by solving $\mathrm{erf}\Big(\dfrac{z_{\alpha/2}}{\sqrt{2}}\Big) = 1 - \alpha$
3. Compute $\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}$
4. Compute $T_\alpha = \left[\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}\ \ , \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}\right]$

Interpretation the result

- If T_α is entirely negative, it suggest a negative difference at population level
- If T_α is entirely positive, it suggest a positive difference at population level
- If T_α is contains zero, it suggest possibly no difference at population level

10.5 Confidence interval for difference of normal means with unknown variance

We use the same formula to *approximate* the CI.

What about using student-t distribution? Nah that is too much for this course. Go study yourself.

Example (stock) Two group of values are provided

$$\begin{aligned}\boldsymbol{y}_A = (y_1^A, y_2^A, ..., y_7^A) = &\quad 34, 28.9, 45.4, 53.2, 29.0, 36.5, 32.9 \\ \boldsymbol{y}_B = (y_1^B, y_2^B, ..., y_8^B) = &\quad 53.2, 33.6, 36.6, 42, 33.3, 37.8, 31.2, 43.4\end{aligned}$$

If we want to know "on average, is there any difference between the two groups", we construct a confidence interval of $\mu_A - \mu_B$.

Solution

1. Compute $\overline{y}_A$ (the maximum likelihood estimator of μ_A)

$$\overline{y}_A = \frac{34 + 28.9 + 45.4 + 53.2 + 29.0 + 36.5 + 32.9}{7} = 37.1286$$

   Compute $\overline{y}_B$ (the maximum likelihood estimator of μ_B)

$$\overline{y}_B = \frac{53.2 + 33.6 + 36.6 + 42 + 33.3 + 37.8 + 31.2 + 43.4}{8} = 38.8875$$

   Compute $\hat{\mu}_A - \hat{\mu}_B$ as $\overline{y}_A - \overline{y}_B = 37.1286 - 38.8875$

2. For $\alpha = 0.05$, compute $z_{\alpha/2}$ by solving $\mathrm{erf}\Big(\frac{z_{\alpha/2}}{\sqrt{2}}\Big) = 1 - \alpha$ gives $z_{\alpha/2} = 1.96$

3. We *estimate* σ_A^2, σ_B^2 by unbiased estimator of variance

$$\begin{aligned}\hat{\sigma}_A^{\text{unbiased}} &= \frac{1}{n_A - 1}\sum_{i=1}^{n_A}(y_i^A - \overline{y}_A)^2 = 81.4257 \\ \hat{\sigma}_B^{\text{unbiased}} &= \frac{1}{n_B - 1}\sum_{i=1}^{n_B}(y_i^B - \overline{y}_B)^2 = 51.3698\end{aligned}$$

   Therefore

$$\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{81.4257}{7} + \frac{51.3698}{8}} = 4.2489$$

4. The 95% confidence interval of $\mu_A - \mu_B$ is

$$[-1.7589 - 1.96(4.2489), \quad -1.7589 - 1.96(4.2489)] = [-10.0867, 6.5689]$$

   As the interval contains zero, we cannot rule out the possibility of there being no difference at a population level.

11 Hypothesis testing

Motivation / background

- Point estimation (estimator), confidence interval and hypothesis testing are all doing the same job: telling something from an observed dataset
  - Point estimation / estimator: output a single number to describe the dataset (such as estimator of the mean)
  - Confidence interval: give a range of plausible values for the unknown population parameter
  - Hypothesis testing: gives the probability that the data satisfy certain hypothesis

11.1 Null hypothesis and alternative hypothesis

- In a crime scene, you are a detective, you collect information (data) and your job is to prove someone is a murderer (alternative hypothesis).
- We take the null hypothesis (innocent) as the default position: "Presumed innocent until proven guilty"
- Mathematically, we write

$$\begin{aligned} H_0 &: \text{ null hypothesis} \\ H_A &: \text{ alternative hypothesis} \end{aligned}$$

- What we are doing here: based on the observed data $\boldsymbol{y}$, we *ask how much evidence the data carries* **against** *the null hypothesis*
  - This is called Neyman-Pearson theory of statistical testing
  - We are asking "what is the probability of seeing $\boldsymbol{y}$ given the null hypothesis is true"
  - The smaller is this probability, the stronger the evidence against null hypothesis

  In the detective story, we have

Hypothesis	In English	In English (simplified)	Conditional probability	Conditional probability
H_0	He is innocent	He is good	$\mathbb{P}(\boldsymbol{y} \,\vert\, \text{he is good})$	$\mathbb{P}(\boldsymbol{y} \,\vert\, H_0)$
H_A	He is the murderer	He is bad	$\mathbb{P}(\boldsymbol{y} \,\vert\, \text{he is bad})$	$\mathbb{P}(\boldsymbol{y} \,\vert\, H_A)$

  What is hypothesis testing = find the value of the probability $\mathbb{P}(\boldsymbol{y} \,|\, \text{he is good})$
  - This probability $\mathbb{P}(\boldsymbol{y} \,|\, \text{he is good})$ is known as the **p-value**
  - If $\mathbb{P}(\boldsymbol{y} \,|\, \text{he is good})$ is small $\iff$ improbable $(\boldsymbol{y} \,|\, \text{he is good})$ occur $\Longleftarrow$ he is bad $\iff$ H_A is true
  - If $\mathbb{P}(\boldsymbol{y} \,|\, \text{he is good})$ is large $\iff$ likely $(\boldsymbol{y} \,|\, \text{he is good})$ occur $\Longleftarrow$ he is good $\iff$ H_0 is true
  - It is important to note the direction of the arrow $\Longleftarrow$ and understand what it means
    - $A \Longleftarrow B$ means "if B then A", it says something of A from B
    - $A \Longleftarrow B$ says nothing of B from A
  - In hypothesis testing,
    - small p-value means we have lots of evidence against the null
    - large p-value means we have little evidence against the null
      - "little evidence against the null" does not mean null is true
      - we can only say "it is inconclusive" / "no conclusion"
      - why we have to say "this is inconclusive": possibly due to small sample size

11.2 p-value of testing normal mean with known variance: two-sided test

- We are given dataset $\boldsymbol{y} = (y_1, ..., y_n)$
- We believe data are drawn from $Y \sim \mathcal{N}(\mu, \sigma^2)$
- We assume σ^2 known and μ unknown
- We have a guess μ_{guess}, we want to test whether $\mu = \mu_{\text{guess}}$ is true

$$\begin{aligned} H_0 &: \mu = \mu_{\text{guess}} \\ H_A &: \mu \neq \mu_{\text{guess}} \end{aligned}$$

- What we do: we get the information ($\hat{\mu}$ here) and ask "how unlikely is the estimate $\hat{\mu}$ we have observed if the population mean was $\mu = \mu_{\text{guess}}$?
- In other words, we are asking

$$\text{How small is the probability } \mathbb{P}(\hat{\mu} \mid \mu = \mu_{\text{guess}})$$

- We use maximum likelihood estimator (i.e., sample mean) for the mean

$$\hat{\mu} = \overline{y} = \text{average}(y_1, y_2, \ldots, y_n)$$

– Because we assumed data are drawn from $Y \sim \mathcal{N}(\mu, \sigma^2)$, so by the finite-sample statistics of sample mean,

$$\overline{y} \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big)$$

– The null hypothesis H_0 is that $\mu = \mu_{\text{guess}}$, hence **finite-sample statistics of sample mean if H_0 is true** gives

$$\overline{y} \sim \mathcal{N}\Big(\mu_{\text{guess}}, \frac{\sigma^2}{n}\Big)$$

By the fact that $X \sim \mathcal{N}(\mu, \sigma^2) \iff \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, we have

$$\overline{y} \sim \mathcal{N}\Big(\mu_{\text{guess}}, \frac{\sigma^2}{n}\Big) \iff \frac{\overline{y}-\mu_{\text{guess}}}{\sqrt{\dfrac{\sigma^2}{n}}} = \frac{\overline{y}-\mu_{\text{guess}}}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{\#}$$

– Now let $z_{\overline{y}} = \dfrac{\overline{y}-\mu_{\text{guess}}}{\sigma/\sqrt{n}}$. As we are looking for evidence **against the null**, we look for the probability

$$\mathbb{P}\bigg(\text{NOT}\Big\{\overline{y} \sim \mathcal{N}\Big(\mu_{\text{guess}}, \frac{\sigma^2}{n}\Big)\Big\}\bigg) \overset{(\#)}{=} \mathbb{P}\bigg(\text{NOT}\Big\{\frac{\overline{y}-\mu_{\text{guess}}}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)\Big\}\bigg)$$
$$\overset{z_{\overline{y}}=\frac{\overline{y}-\mu_{\text{guess}}}{\sigma/\sqrt{n}}}{=} \mathbb{P}\bigg(\text{NOT}\Big\{z_{\overline{y}} \sim \mathcal{N}(0,1)\Big\}\bigg)$$

– By complementary rule in probability

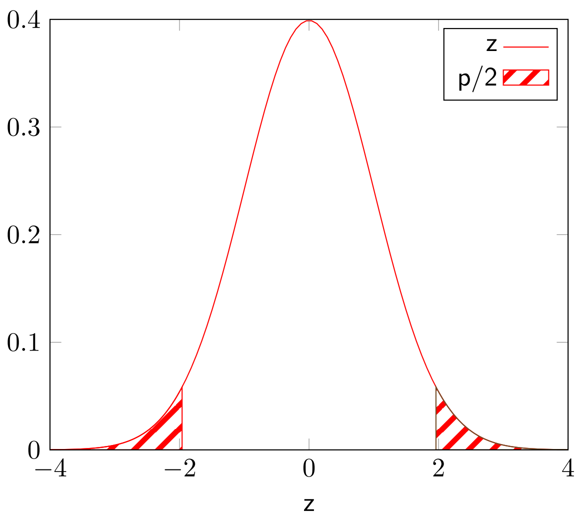
$$\begin{aligned}
\mathbb{P}\bigg(\text{NOT}\Big\{z_{\overline{y}} \sim \mathcal{N}(0,1)\Big\}\bigg) &= 1 - \mathbb{P}\Big(z_{\overline{y}} \sim \mathcal{N}(0,1)\Big) && \mathbb{P}(\text{not } E) = 1 - \mathbb{P}(E) \\
&= 1 - \mathbb{P}\big(-|z_{\overline{y}}| \leq Z \leq |z_{\overline{y}}|\big) \\
&= \mathbb{P}\big(Z \leq -|z_{\overline{y}}|\big) + \mathbb{P}\big(Z \geq |z_{\overline{y}}|\big) \\
&= 2\mathbb{P}\big(Z \leq -|z_{\overline{y}}|\big) && \text{normal distribution is symmetric}
\end{aligned}$$

– p-value in this case is defined as

$$p = 2\mathbb{P}\big(Z \leq -|z_{\overline{y}}|\big)$$

  * In this case we are asking how close are $\overline{y}, \mu_{\text{guess}}$ to each other, measured as $|\overline{y} - \mu_{\text{guess}}|$.
    · We are using absolute value here, so it doesn't matter $\begin{cases} \overline{y} \text{ is larger than } \mu_{\text{guess}} \\ \overline{y} \text{ is smaller than } \mu_{\text{guess}} \end{cases}$
  * if $p < 0.01$, we have strong evidence against the null
  * otherwise, we have no conclusion.

– Graphically, we are doing the following
  * We turn the null hypothesis into a standard normal distribution variable $z_{\overline{y}}$
  * If H_0 is true, then $z_{\overline{y}}$ is likely to be within the standard normal distribution $\mathcal{N}(0,1)$
  * We look for the probability that $z_{\overline{y}}$ is NOT within $\mathcal{N}(0,1)$,
    i.e., we look for the area of the two tails of the z-curve
  * These are represent how likely $z_{\overline{y}}$ is NOT within $\mathcal{N}(0,1)$



11.3 p-value of testing normal mean with known variance: one-sided test

- We are given dataset $\boldsymbol{y} = (y_1, ..., y_n)$
- We believe data are drawn from $Y \sim \mathcal{N}(\mu, \sigma^2)$
- We assume σ^2 known and μ unknown
- We have a guess μ_{guess}, we want to test whether $\mu \leq \mu_{\text{guess}}$ is true

$$\begin{array}{rcl} H_0 & : & \mu \leq \mu_{\text{guess}} \\ H_A & : & \mu > \mu_{\text{guess}} \end{array}$$

- Under similar analysis as before, this time we look for $p = \mathbb{P}(Z > z_{\overline{y}})$, where $z_{\overline{y}} = \dfrac{\overline{y} - \mu_{\text{guess}}}{\sigma/\sqrt{n}}$.

11.4 The p-value for hypothesis testing of normal mean with known variance

- We are given dataset $\boldsymbol{y} = (y_1, ..., y_n)$
- We believe data are drawn from $Y \sim \mathcal{N}(\mu, \sigma^2)$
- We assume σ^2 known and μ unknown
- We have a guess μ_{guess}
- The steps for calculating p-values are
  1. Calculate the maximum likelihood estimator of mean / sample mean $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$
  2. Perform standardization to get the standard z-score $z_{\overline{y}} = \dfrac{\overline{y} - \mu_{\text{guess}}}{\sigma/\sqrt{n}}$
  3. Calculate the p-value:

$$p = \begin{cases} 2\mathbb{P}(Z - |z_{\overline{y}}|) & H_0 : \mu = \mu_{\text{guess}} \text{ vs } H_A : \mu \neq \mu_{\text{guess}} \\ 1 - \mathbb{P}(Z < z_{\overline{y}}) & H_0 : \mu \leq \mu_{\text{guess}} \text{ vs } H_A : \mu > \mu_{\text{guess}} \\ \mathbb{P}(Z < z_{\overline{y}}) & H_0 : \mu \geq \mu_{\text{guess}} \text{ vs } H_A : \mu < \mu_{\text{guess}} \end{cases}$$

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

11.5 The p-value for hypothesis testing of normal mean with unknown variance

- We are given dataset $\boldsymbol{y} = (y_1, ..., y_n)$
- We believe data are drawn from $Y \sim \mathcal{N}(\mu, \sigma^2)$
- We assume both σ^2 and μ unknown
- We have a guess μ_{guess}
- The steps for calculating p-values are
  1. Calculate the maximum likelihood estimator of mean / sample mean $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$
  2. Calculate the unbiased estimator of variance

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

  3. Perform standardization to get the t-score $t_{\overline{y}} = \dfrac{\overline{y} - \mu_{\text{guess}}}{hat\sigma_{\text{unbiased}}/\sqrt{n}}$
  4. Calculate the p-value:

$$p = \begin{cases} 2\mathbb{P}(T - |t_{\overline{y}}|) & H_0 : \mu = \mu_{\text{guess}} \text{ vs } H_A : \mu \neq \mu_{\text{guess}} \\ 1 - \mathbb{P}(T < t_{\overline{y}}) & H_0 : \mu \leq \mu_{\text{guess}} \text{ vs } H_A : \mu > \mu_{\text{guess}} \\ \mathbb{P}(T < t_{\overline{y}}) & H_0 : \mu \geq \mu_{\text{guess}} \text{ vs } H_A : \mu < \mu_{\text{guess}} \end{cases}$$

  where $T \sim T(n-1)$ is the standard student-t distribution with degree-of-freedom $n-1$.

11.6 The p-value for hypothesis testing of difference of normal mean, known variance

Two-sided test

- The hypothesis testing is
$$\begin{array}{rcl} H_0 & : & \mu_A = \mu_B \\ & \text{vs} & \\ H_A & : & \mu_A \neq \mu_B \end{array}$$
- If null is true
$$\overline{y}_A - \overline{y}_B \sim \mathcal{N}\Big(0,\ \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\Big)$$
- The z-score
$$z_{\overline{y}_A - \overline{y}_B} = \frac{\overline{y}_A - \overline{y}_B}{\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}}$$
- The p-value
$$p = 2\mathbb{P}\Big(Z < -|z_{\overline{y}_A - \overline{y}_B}|\Big)$$

One-sided test

- The hypothesis testing is
$$\begin{array}{rcl} H_0 & : & \mu_A \geq \mu_B \\ & \text{vs} & \\ H_A & : & \mu_A < \mu_B \end{array}$$
Then
$$p = \mathbb{P}\Big(Z < z_{\overline{y}_A - \overline{y}_B}\Big)$$

11.7 The p-value for hypothesis testing of difference of normal mean, unknown variance

Nah too complicated for this course.