Lecture 1, Jan 9, 2023

Uncertainty

- Sources:
  - Limited measurements – more measurements = less uncertainty, e.g. processes that are difficult to measure
  - Difficult-to-model scenarios, e.g. weather, dice roll
- Probability allows us to be systematic in the face of uncertainty ("be consistent about what we don't know")

The Coin Flip

- Heads or tails
- A fair coin has $P(H) = 0.5 = P(T)$; unbalanced coins have different probabilities, but we must have $P(H) + P(T) = 1$ because *something* has to happen 100% of the time
- Example: $P(H) = 0.3, P(T) = 0.7$, then:
  - $P(HH) = P(H)P(H) = 0.09$
  - $P(HT) = P(H)P(T) = 0.21 = P(T)P(H) = P(TH)$
    - This is only true because of *independence*, as the two flips are uncorrelated
  - $P(HT \text{ or } TH) = P(HT) + P(TH) = 0.42$
  - $P(HT, TH, HH \text{ or } TT) = 1$
  - We can construct a fair coin by setting "heads" to HT and "tails" to TH and ignoring all other outcomes

Core Concepts

Definition

The *sample space* S is the set of all possible outcomes

- e.g. for a single coin flip, $S = \{ H, T \}$; for two coin flips, $S = \{ TH, TT, HH, HT \}$

Definition

An *event* is a subset of a sample space, i.e. some subset of possible outcomes

The *complement* of an event A is $A' = \{ a \mid a \in S, a \notin A \}$, the set of everything in S that is not in A

- e.g. $\{ 1, 2, 3, 4, 5, 6 \}$ or $\{ \text{even}, \text{odd} \}$ could both be the set of all events for a die
- Example: $S = \left\{ (x, y) \;\middle|\; x^2 + y^2 \leq 1 \right\}$ (the unit circle), $A = \{ (x, y) \mid (x, y) \in S, x \geq 0 \}$, then $A' = \{ (x, y) \mid (x, y) \in S, x < 0 \}$

Definition

The *intersection* of 2 events A and B, $A \cap B$, is everything in S that is in both A and B

The *union* of 2 events $A \cup B$ is everything in A or B

- e.g. for a die: $\{ \text{even} \} \cap \{ n \leq 3 \} = \{ 2 \}$
- $A \cap A' = \varnothing, A \cup A' = S$

Lecture 2, Jan 11, 2023

Counting

- Counting is the problem of finding the number of elements in some event A
  - e.g. for a coin flip, if $A = \{ H \}$, then we have one element; for a die, if $A = \{ \text{even} \}$, then we have 3 elements
- Two events A and B are *mutually exclusive* if $A \cap B = \varnothing$
  - For two events that are mutually exclusive, we can add up their number of elements when counting

Multiplying Options

- Where we can choose 1 option from each category, we multiply the category sizes together
  - e.g. choosing a president and VP from n people has $n(n-1)$ possibilities
- Example: How many even 4-digit numbers can we make from $\{ 0, 1, 2, 5, 6, 9 \}$?
  - Consider events A and B:
    * In A, 0 is the last digit so A has $1 \cdot 5 \cdot 4 \cdot 3 = 60$ elements
    * In B, 0 is not the last digit (A and B are mutually exclusive); the last digit could be 2 or 6 and the first digit can be anything but zero or what we chose for the last digit, so we have $2 \cdot 4 \cdot 4 \cdot 3 = 96$ elements
    * Since $A \cap B = \varnothing$, the total count is 156

Permutations

- A *permutation* is an ordering of the elements in an event
- Given n items, there are $n!$ permutations
- If we want to permute r items out of n, there are $\dfrac{n!}{(n-r)!}$ permutations
- With n slots to fill, where there are m kinds of items and n_k of each item, the number of permutations is $\dbinom{n}{n_1, n_2, \cdots, n_m} = \dfrac{n!}{n_1!n_2!\cdots n_m!}$
- Example: How many distinct ways can we order "ATLANTIC"?
  - $\dbinom{8}{2, 2, 1, 1, 1, 1} = \dfrac{8!}{2!2!1!1!1!1!} = 10080$
- Example: If we flip a coin 10 times, how many sequences have 4 heads?
  - We're looking at combinations of $HHHHTTTTTT$, so $\dbinom{n = 10}{n_1 = 4, n_2 = 6} = \dfrac{10!}{4!6!} = 210$
  - This gives us a probability of getting 4 heads of $\dfrac{210}{2^10} = \dfrac{210}{1024}$ (assuming a fair coin)

Lecture 3, Jan 13, 2023

Partitions

Definition

A *partition* is a way to divide a number of elements into groups of certain sizes, where we don't care about the order of elements in the groups

The number of partitions of n distinct objects with m bins of size $n_1, \cdots, n_m$ is

$$\binom{n}{n_1, n_2, \cdots, n_m} = \frac{n!}{n_1!n_2!\cdots n_m!}$$

- Example: Partitions of $\{ a, b, c \}$ with $n_1 = 2, n_2 = 1$

  - Possible partitions are $\{ab, c\}, \{ac, b\}$ or $\{bc, a\}$, so there are 3 partitions
  - Notice $3 = \frac{3!}{2!1!}$
- Formally, let there be n distinct objects and m bins which can hold $n_1, n_2, \cdots, n_m$ objects such that $\sum_{k=1}^{m} n_k = n$, then the number of partitions is $\binom{n}{n_1, n_2, \cdots, n_m} = \frac{n!}{n_1!n_2!\cdots n_m!}$
  - This is the same calculation as permutations with identical objects
  - If $n_k = 1$, then the number of partitions is $n!$ as we just have a permutation; then we divide through by the number of possible permutations within each box since we don't care about the order of elements within boxes

Combinations

- A combination is a permutation where we don't care about the order
- Given n objects, the number of subsets of size r we can make is $\binom{n}{r} = \binom{r}{r, n-r} = \frac{n!}{(n-r)!r!}$
  - This is known as "n choose r"
- In terms of partitions, we have one box of size r (the elements we're choosing), and another box of size $n - r$ (the elements we're not choosing)

Probability

- Given a sample space S and an event $A \subseteq S$, then the probability of event A is $P(A) \in [0, 1]$
  - $P(S) = 1$
  - If $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$
- Example: What's the probability of getting 2 aces and 3 jacks in a hand?
  - Ways to get 2 aces: $\binom{4}{2} = 6$
    - * Out of 4 aces in the deck we're picking 2 of them
  - Ways to get 3 jacks: $\binom{4}{3} = 4$
  - Therefore we have 24 such hands, so the probability is $\frac{24}{\binom{52}{5}} = 0.9 \times 10^{-5}$

Lecture 4, Jan 16, 2023

Probability of Unions

Equation

The probability of the union of two events $P(A \cup B)$ is

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

- In general $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - $P(A \cup B)$ is the overlap of A and B, which is counted twice by $P(A) + P(B)$
- Example: Probability of rolling a 7 or having at least one 2
  - Probability of rolling a 7 is $\frac{1}{6}$
  - Probability of rolling at least one 2 is $\frac{11}{36}$
  - Probability of rolling a 7 and having at least one 2 is $\frac{2}{36}$ (2 and 5, or 5 and 2)
  - Final probability is $\frac{15}{36} = \frac{5}{12}$

- For the union of more than 2 events, we can expand the rule:
  - $P(A \cup B \cup C) = P(A) + P(B \cup C) + P(A \cap (B \cup C))$
    $= P(A) + P(B) + P(C) - P(B \cap C) - P((A \cap B) \cup (A \cap C))$
    $= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)$
  - Note intersection distributes over union

Conditional Probability

Definition

If A and B are both events, the probability of B *conditioned* on A (or probability of B given A), $P(B|A)$ is the probability of B given that A occurs

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

- This formula arose due to $P(A \cap B) = P(A)P(B|A)$, i.e. the probability of both A and B happening is the probability of A happening times the probability of B happening given that A happens

Lecture 5, Jan 18, 2023

Independence

Definition

Two events A and B are *independent* if $P(A|B) = P(A)$ or $P(B|A) = P(B)$

- Note, $P(B|A) = P(B) = \dfrac{P(A \cap B)}{P(A)} \implies P(A \cap B) = P(A)P(B) \implies P(A|B) = \dfrac{P(A)P(B)}{P(B)} = P(A)$
- i.e. Two events are independent if one of them happening does not affect the probability of the other
- Note: Independence is not the same as mutual exclusion! e.g. for a coin flip, heads and tails are mutually exclusive, but they are not independent since $P(H|T) = 0 \neq P(H)$

Bayes' Rule

- Suppose $P(A) > 0, P(B) > 0$; we know $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$ so $P(A|B)P(B) = P(B|A)P(A)$
- Rearrange to get $\dfrac{P(A|B)}{P(A)} = \dfrac{P(B|A)}{P(B)}$

Definition

Bayes' Rule:

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

or

$$P(A|B)P(B) = P(B|A)P(A)$$

- Bayes' rule is useful for reversing causality; based on what we saw, we can make inferences about what caused it

Partitions and Probability

- Recall $B_1, \cdots, B_k$ is a partition if $B_i \cap B_j = \varnothing$ whenever $i \neq j$ and $B_1 \cup B_2 \cdots \cup B_k = S$

Equation

The law of total probability: Given a partition $B_1, \cdots, B_k$ and some event A,

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \end{aligned}$$

- Proof:
  - Observe $A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k)$
  - Then $P(A) = P((A \cap B_1) \cup \cdots \cup (A \cap B_k))$
  - Note when we partitioned the sample space, each B_i is mutually exclusive, therefore $A \cap B_i$ are also mutually exclusive
  - Therefore $P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k)$ by mutual exclusivity
  - $P(A) = \displaystyle\sum_{i=1}^{k} P(A \cap B_i)$
- Example: Machines 1, 2, 3 make products 30%, 45%, and 25% of the time respectively (as a partition); the product is defective 2%, 3%, and 2% of the time for these machines respectively, what is the probability that the product is defective
  - $P(\text{defective}) = P(\text{defective}|\text{made by 1})P(\text{made by 1}) + P(\text{defective}|\text{made by 2})P(\text{made by 2}) + P(\text{defective}|\text{made by 3})P(\text{made by 3})$
  - $P(\text{defective}) = 0.30 \cdot 0.02 + 0.45 \cdot 0.03 + 0.25 \cdot 0.02 = 2.45\%$
- In general independence, partitions, and conditional probability show up in the problem statement and it's up to us to interpret them

Lecture 6, Jan 20, 2023

Bayes' Rule and Total Probability

- Given a partition of A into $C_1, \cdots, C_k$
- $\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(A)} \\ &= \frac{P(B)P(A|B)}{\sum_{i=1}^{k} P(A \cap C_i)} \\ &= \frac{P(B)P(A|B)}{\sum_{i=1}^{k} P(A|C_i)P(C_i)} \end{aligned}$
  - Often $B = C_n$ for some $n = 1, \cdots, k$

Random Variables

Definition

A *random variable* is a function that maps each element of a sample space to a real number

- We denote a random variable with capital letters, e.g. X, Y
- In the discrete case, the random variable can only take on a finite (or countably infinite) set of values

- In the continuous case the random variable can take any value in the real numbers
- We write $X = x$, with the lowercase x to denote values that the random variable can take on
- Example: coin flips
  - $S = \{ H, T \}$; our random variable can be $X = \begin{cases} 0 & H \\ 10 & T \end{cases}$
  - If we do 3 coin flips, X can be the number of heads

Lecture 7, Jan 23, 2023

Probabilities of Discrete Random Variables

Definition

$f(x)$ is the *probability mass function* (PMF) of a discrete random variable X if:

- $f(x) = P(X = x), \forall x \in S$
- $\sum\limits_{x \in S} f(x) = 1$
- $f(x) \geq 0, \forall x \in S$

Definition

The *cumulative distribution function* (CDF) of a discrete random variable X with PMF $f(x)$ is

$$F(x) = \sum_{t \leq x} f(t)$$

or equivalently

$$F(x) = P(X \leq x)$$

- e.g. flipping a coin 3 times:
  - $f(x) = \begin{cases} \frac{1}{8} & x = 0, x = 3 \\ \frac{3}{8} & x = 1, x = 2 \end{cases}$
  - $F(-1) = 0, F(0) = \frac{1}{8}, F(1) = \frac{1}{2}, F(2) = \frac{7}{8}, F(3) = 1$
  - $P(X \geq 2) = 1 - P(X \leq 1) = 1 - F(1) = \frac{1}{2}$
- Properties of CDFs:
  - $F(-\infty) = 0$
  - $F(\infty) = 1$
  - All CDFs are nondecreasing

Probabilities of Continuous Random Variables

- In the continuous case, the probability of the random variable equalling any specific value is zero, since there are an uncountably infinite number of outcomes in every interval
- For a continuous random variable we can only talk about probabilities of the variable being in some interval

Definition

$f(x)$ is the *probability distribution/density function* (PDF) of a continuous random variable X if:

- $\displaystyle\int_a^b f(x)\,\mathrm{d}x = P(a \leq X \leq B), \forall a \leq b \in \mathbb{R}$
- $\displaystyle\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
- $f(x) \geq 0, \forall x \in S$

Definition

The *cumulative distribution function* (CDF) of a continuous random varaible X with PDF $f(x)$ is

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

or equivalently

$$F(x) = P(X \leq x)$$

- This gives $P(a \leq X \leq b) = F(b) - F(a)$
- The properties of CDFs from the discrete case carry over

Lecture 8, Jan 25, 2023

Joint Probability Distributions (Discrete)

- Often events are correlated, and we want to look at the probability of two events together

Definition

$f(x, y)$ is the *joint probability mass function* of the discrete varaibles X and Y if:

- $f(x, y) \geq 0, \forall (x, y) \in S$
- $\displaystyle\sum_{(x,y)\in S} f(x, y) = 1$
- $f(x, y) = P(X = x, Y = y)$

- Example: Drawing 5 cards from a deck of 52, X is the number of queens and Y is the number of kings, what is $f(x, y)$?
  - Total number of hands is $\displaystyle\binom{52}{5}$
  - Number of hands with x queens and y kings is $\displaystyle\binom{4}{x}\binom{4}{y}\binom{52-8}{5-x-y}$ for $x, y \leq 4, x + y \leq 5$
  - Therefore $f(x, y) = \begin{cases} 0 & x + y > 5, x > 4, y > 4 \\ \dfrac{\binom{4}{x}\binom{4}{y}\binom{44}{5-x-y}}{\binom{52}{5}} & x + y \leq 5, x \leq 4, y \leq 4 \end{cases}$
- Continued example: Consider $A = \{\,(x + y) \mid x + y = 2\,\}$; then $\displaystyle P((x, y) \in A) = \sum_{(x,y)\in A} f(x, y) = f(0, 2) + f(1, 1) + f(2, 0)$

Joint Probability Distributions (Continuous)

Definition

$f(x,y)$ is the *joint probability density function* of the continuous variables X and Y if

- $f(x,y) \geq 0, \forall (x,y) \in S$
- $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$
- If $A \in S$, $\displaystyle P((x,y) \in A) = \int_{(x,y)\in A} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

- Example: Uniform distribution, $S = \{\,(x,y) \mid 1 \leq x \leq 1, -1 \leq y \leq 1\,\}$
  - A uniform distribution means the probability of any outcome is the same
  - $f(x,y) = \dfrac{1}{4}$ for $(x,y) \in S$, since the "area" is 4
  - Consider an event $A = (x,y)|0 \leq x \leq 1, 0 \leq y \leq 1$, now $P((x,y) \in A) = \dfrac{1}{4}$ since A takes up a quarter of S
    - * We can also use $\displaystyle\int_0^1\int_0^1 \frac{1}{4}\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{4}$

Marginal Distributions

- Suppose we know $f(x,y)$ for X, Y; we can find the distribution for X as $\displaystyle g(x) = \sum_y f(x,y)$ in the discrete case, and $\displaystyle g(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$ in the continuous case
  - This is known as a *marginal distribution*
- Example: Given a PDF $f(x,y) = \begin{cases} 1 \\ |x| + |y| \leq 1, y \geq 0 \end{cases}$ 0otherwise, find $g(x)$
  - This describes a triangle above the x axis
  - $\displaystyle g(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \int_0^{1-|x|} 1\,\mathrm{d}y = 1 - |x|$

Lecture 9, Jan 30, 2023

Independence of Random Variables

Definition

Let X and Y be two random variables with joint probability distribution $f(x,y)$ and marginal distributions $g(x), h(y)$; these random variables are *independent* if $f(x,y) = g(x)h(y)$

- This definition applies to both continuous and discrete cases

Expectation

Definition

Let X be a random variable with distribution $f(x)$; the expectation value of X is, in the discrete case:

$$E[x] = \sum_x x f(x)$$

in the continuous case:

$$E[x] = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x$$

- Also known as expected value or the mean, denoted by μ
- We may replace x with a function of the random variable $g(x)$ to find its expectation

Definition

Let X and Y be random variables with joint distribution function $f(x, y)$; the expectation value of the function $g(X, Y)$ is, in the discrete case:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y)$$

in the continuous case:

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

- Expectation values easily generalize to multiple variables

Lecture 10, Feb 1, 2023

Variance

Definition

Let X be a random variable with distribution $f(x)$, then the *variance* of X is

$$\sigma^2 = \operatorname{var}(X) = E[(X - \mu)^2]$$

which is in the discrete case:

$$\sum_x (x - \mu)^2 f(x)$$

in the continuous case:

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, \mathrm{d}x$$

$\sigma = \sqrt{\sigma^2} = \sqrt{\operatorname{var}(X)}$ is known as the *standard deviation* of X

- Variance is a measure of variability, or spread – how wide a range we're going to see values from a distribution
- Example: uniform distribution $f(x) = \begin{cases} \frac{1}{a} & 0 \leq x \leq a \\ 0 & \text{elsewhere} \end{cases}$

  - $\sigma^2 = \int_0^a \left(x - \frac{a}{2}\right)^2 \frac{1}{a}\,\mathrm{d}x = \frac{1}{a}\left[\frac{x^3}{a} - a\frac{x^2}{2} + a^2\frac{x^2}{4}\right]_0^a = \frac{a^2}{12}$
  - Note $\sigma^2 \to 0$ as $a \to 0$, which would give us a delta function
- $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$

  $= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)\,\mathrm{d}x$

  $= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - 2\mu \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + \mu^2 \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$

  $= E[X^2] - 2\mu^2 + \mu^2$

  $= E[X^2] - \mu^2$

  $= E[X^2] - (E[X])^2$
  - This also applies to the discrete case since sums can be split exactly in the same way

Covariance and Correlation

Definition

Let X and Y be random variables with joint distribution $f(x, y)$ and means μ_x, μ_y, then the *covariance* of X and Y is

$$\sigma_{xy} = \mathrm{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

which is in the discrete case

$$\sum_x \sum_y (x - \mu_x)(y - \mu_y) f(x, y)$$

in the continuous case

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

- Note that $\sigma_{xx} = E[(X - \mu_x)(X - \mu_x)] = E[(X - \mu_x)^2] = \sigma^2$, i.e. the variance is the covariance of a random variable with itself
- The covariance is a measure of correlation
  - If the covariance is positive, then both variables tend to be above their means or below their means at the same time; the two variables would be *positively correlated*
  - If the covariance is negative, then when one is above its mean the other would tend to be below its mean; the two variables would be *negatively correlated*
- $\sigma_{xy} = E[XY] - \mu_x\mu_y$

Definition

Let X and Y be random varaibles with covariance σ_{xy} and standard deviations σ_x and σ_y, then the *correlation coefficient* of X and Y is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

- The correlation coefficient is a normalized, dimensionless version of the covariance; we always have $\rho_{xy} \in [-1, 1]$
  - Two variables are *uncorrelated* if $\rho_{xy} = 0$
  - Note independence implies correlation 0, but correlation 0 does not imply independence

Lecture 11 (Recorded)

Expectations of Linear Combinations of Random Variables

Definition

A function $p(x)$ is linear if $p(ax+y) = ap(x) + p(y)$

- Because integration and summation are linear, expectation $E[X]$ is linear: $E[aX+Y] = aE[X] + E[Y]$
- Useful implications:
  - $E[aX+b] = aE[X] + b$
  - $E[g(X,Y) + h(X,Y)] = E[g(X,Y)] + E[h(X,Y)]$

Variance and Independence

- If X and Y are independent: $$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyg(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} xg(x)\,\mathrm{d}x \int_{-\infty}^{\infty} yh(y)\,\mathrm{d}y \\ &= E[X]E[Y] \end{aligned}$$
  - Therefore because $\sigma_{XY} = E[XY] - E[X]E[Y]$, this means independence implies uncorrelated
  - But uncorrelated does not imply independence!
- Note $\sigma^2_{aX+bY+c} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma^2_{XY}$
  - The constant does not affect the variance
  - There is a cross term
  - The cross term σ_{XY} disappears if X and Y are independent variables

Uniform Distribution

- Every element in the sample space has the same probability
- For $S = 1, \cdots, n$, then $f(k) = \dfrac{1}{n}$ for $k \in S$

Binomial Distribution

- A Bernoulli random variable has 2 outcomes $S = 0, 1$, and each has a probability (e.g. a coin flip)

Definition

A *Bernoulli process* is a process involving n repeated, independent, identical trials where the only outcomes possible are 1 and 0, with $P(1) = p$; let X be the number of 1's that occur, then the *Binomial distribution* is the probability mass function for X:

$$P(X = x) = f(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Example: when sending a string of 0s and 1s, how many errors will occur?
- $p^x(1-p)^{n-x}$ is the probability of having x 1's and $n-x$ 0's in some order; $\binom{n}{x}$ is all the ways to have that many 1's and 0's
- $E[X] = np$ is the expectation value of a binomial distribution

    - We can think of the Bernoulli process as a sum of n trials, $X = Y_1 + \cdots + Y_n$, so since expectation is linear, we can just add up the probabilities
- Similarly $\sigma_X^2 = np(1-p)$

Multinomial Distribution

- An extension of the binomial distribution where each trial can have m outcomes instead of just 2
- The chance of each outcome E_i is p_i
- x_i is the number of times we get outcome E_i; $\displaystyle\sum_i x_i = n$

Definition

The *multinomial distribution* is

$$f(x_1, \cdots, x_n; p_1, \cdots, p_m, n) = \binom{n}{x_1, x_2, \cdots, x_m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

Lecture 12, Feb 6, 2023

Hypergeometric Distribution

Definition

Given N total objects, where K of the N are successes, and sampling n times without replacement, the *hypergeometric distribution* describes the probability of x successes and $n - x$ failures:

$$h(x; N, n, K) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

- Statistics:
    - $\mu = \dfrac{nK}{N}$
    - $\sigma^2 = \dfrac{Kn(N-n)}{N(N-1)}\left(1 - \dfrac{K}{N}\right)$

Negative Binomial Distribution and Geometric Distribution

Definition

Given repeated trials with probability p of success and $1-p$ of failure, the *negative binomial distribution* describes the probability of observing the k-th success on trial number x:

$$b^*(x, k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

- i.e. x is how many trials it takes to get k successes
- Intuition:
    - If we are on trial x< then the chance of $k-1$ successes in the last $x-1$ trials is $b(k-1; x-1, p) = \binom{x-1}{k-1} p^{k-1}(1-p)^{x-1-k+1}$
    - Getting the next success is just multiplication by p: $b^*(x; k, p) = pb(k-1; x-1, p) = p\binom{x-1}{k-1} p^{k-1}(1-p)^{x-1-k+1} = \binom{x-1}{k-1} p^k (1-p)^{x-k}$

- The *geometric distribution* is just the negative binomial distribution with $k = 1$

Definition

Given repeated trials with probability p of success, the *geometric distribution* describes the probability of the first success occurring on trial x:

$$g(x; p) = b^*(x; 1, p) = p(1-p)^{x-1}$$

- Statistics of the geometric distribution:
    - $\mu = \dfrac{1}{p}$
    - $\sigma^2 = \dfrac{(1-p)}{p^2}$

Lecture 13, Feb 8, 2022

Poisson Distribution

- Given a sequence of independent intervals, how many times will a given event occur within an interval?
- Example: number of goals in a soccer game, number of snow days in a year, number of arrivals per minute at an internet router
- A *Poisson process* is a process where the number of arrivals in an interval is a random variable independent of the other intervals
    - The number of arrivals is proportional to the length of the interval
    - The number of arrivals in a given interval follows a Poisson PMF

Definition

The *Poisson PMF* is given by

$$p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \cdots$$

where λ is the Poisson parameter and x is the number of arrivals in the interval described by the distribution

The Possion parameter λ is both the mean and the variance of the Possion distribution; we have

$$\lambda = rt$$

where r is the rate of arrivals, and t is the length of the interval

- Statistics of the Poisson PMF:
    - $\mu = \sum_{x=0}^{\infty} x\dfrac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\dfrac{\lambda^x}{(x-1)!} = e^{-\lambda}\lambda\sum_{x=1}^{\infty}\dfrac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$
    - We can also show $\sigma^2 = \lambda$ through a similar calculation
- We can think of the Poisson distribution as the binomial distribution, in the limit where $n \to \infty, p \to 0$ and $np \to \lambda$

- $$\lim_{n\to\infty, p\to 0} \binom{n}{x} p^x (1-p)^{n-x} = \lim_{n\to\infty, p\to 0} \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$
$$= \lim_{n\to\infty, p\to 0} \frac{n^x}{x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$
$$= \frac{\lambda^x}{x!} \lim_{n\to\infty, p\to 0} \left(1-\frac{\lambda}{n}\right)^{n-x}$$
$$= \lim_{n\to\infty, p\to 0} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-x}$$
$$= \lim_{n\to\infty, p\to 0} e^{-\lambda} \cdot 1$$
$$= \frac{e^{-\lambda}\lambda^x}{x!}$$
$$= p(x;\lambda)$$

Lecture 14, Feb 10, 2023

The Normal (Gaussian) Distribution

Definition

Given a mean μ and variance σ^2, the *normal distribution* is given by

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

- This gives a symmetric bell centered around the mean μ, with width proportional to σ
  - μ is a translation
  - σ gets larger as the curve gets wider and flatter
  - $\lim\limits_{\sigma\to 0} n(x;\mu,\sigma) = \delta(x-\mu)$
- The Gaussian is important due to the central limit theorem: taking a large number of random variables and taking their average, it will give the normal distribution regardless of the distribution of the individual random variables
- $$\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x\right)^2 = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \,\mathrm{d}y$$
$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2\sigma^2}} \,\mathrm{d}x \,\mathrm{d}y$$
$$= \frac{1}{2\pi\sigma^2} \int_{0}^{2\pi} \int_{0}^{\infty} e^{-\frac{r^2}{2\sigma^2}} r \,\mathrm{d}r \,\mathrm{d}\theta$$
$$= \frac{1}{4\pi\sigma^2} \int_{0}^{2\pi} \int_{0}^{\infty} e^{-\frac{s}{2\sigma^2}} \,\mathrm{d}s \,\mathrm{d}\theta$$
$$= \frac{1}{2\sigma^2} \int_{0}^{\infty} e^{-\frac{s}{2\sigma^2}} \,\mathrm{d}s$$
$$= \frac{1}{2\sigma^2} \left[-2\sigma^2 e^{-\frac{s}{2\sigma^2}}\right]_0^{\infty}$$
$$= 1$$

- $$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} \sigma \, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} \, \mathrm{d}z + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \, \mathrm{d}z \\ &= 0 + \mu \int_{-\infty}^{\infty} n(z; 0, 1) \, \mathrm{d}z \\ &= \mu \end{aligned}$$
  - Substitute $z = \dfrac{x - \mu}{\sigma}$
  - $\displaystyle\int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} \, \mathrm{d}z = 0$ because the integrand is an odd function
- Using a similar argument we may show that the variance is σ^2

Definition

$n(x; 0, 1)$ is referred to as the *standard normal distribution*

$$\Phi(x) = \int_{-\infty}^{x} n(y; 0, 1) \, \mathrm{d}y$$

is the cumulative distribution function of the standard normal, so

$$P(A \leq X \leq B) = \Phi(B) - \Phi(A)$$

- Note Φ is not analytically evaluable, so there are usually tables of values for it
- Suppose X has PDF $n(x; \mu, \sigma)$; let $Z = \dfrac{X - \mu}{\sigma}$, then Z has PDF $n(x; 0, 1)$, which is the standard normal
  - $$\begin{aligned} P(X \leq x) &= \int_{-\infty}^{x} n(x; \mu, \sigma) \, \mathrm{d}x \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{\sigma^2}} \, \mathrm{d}t \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} \sigma \, \mathrm{d}s \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} n(s; 0, 1) \, \mathrm{d}s \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right) \end{aligned}$$
- Therefore $P(A \leq X \leq B) = \Phi\left(\dfrac{A - \mu}{\sigma}\right) - \Phi\left(\dfrac{B - \mu}{\sigma}\right)$
- Example: Suppose X is a random variable with distribution $n(x; 5, 2)$; find $P(-1 \leq X \leq 4)$
  - Need to transform this into the standard normal
  - Let $Z = \dfrac{X - 5}{2}$ then Z has the standard normal distribution and CDF Φ
  - $P(-1 \leq X \leq 4) = P\left(\dfrac{-1 - 5}{2} \leq \dfrac{X - 5}{2} \leq \dfrac{4 - 5}{2}\right) = P\left(-3 \leq Z \leq -\dfrac{1}{2}\right) = \Phi\left(-\dfrac{1}{2}\right) - \Phi(-3) = 0.3072$
  - $\Phi(x)$ is `normcdf` in MATLAB

Lecture 15, Feb 13, 2023

The Normal as a Limit of the Binomial

- Recall that for the binomial distribution $\mu = np, \sigma^2 = np(1-p)$
- Let X be distributed according to the binomial distribution, and let $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$
- As $n \to \infty$, Z approaches the standard normal

The Gamma Distribution

Definition

The Γ function is defined as:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x, \alpha > 0$$

- Note $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$
  - Γ can be thought of as a continuous generalization of the factorial

Definition

The *gamma distribution* with parameters α, β is

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

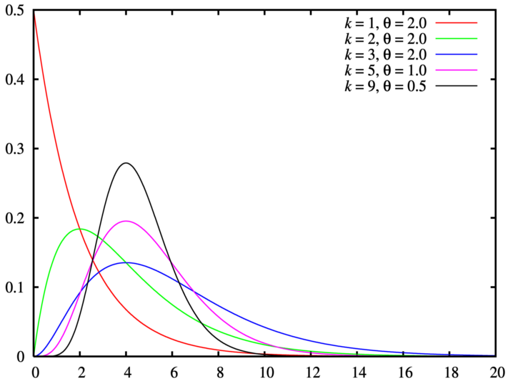

Figure 1: The gamma distribution ($k = \alpha$, $\theta = \beta$)

- Statistics:
  - $\mu = \alpha\beta$
  - $\sigma^2 = \alpha\beta^2$
- The gamma distribution combines and generalizes multiple distributions
- Note $\dfrac{1}{\beta^\alpha \Gamma(\alpha)}$ is just normalization (there is no x); the gamma function has no real effect on the shape of the distribution

The Chi-Squared Distribution

Definition

The χ^2 *distribution* is

$$f(x;v) = f\left(x;\alpha=\frac{v}{2},\beta=2\right) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\frac{v}{2}-1}e^{-\frac{x}{2}} & x>0 \\ 0 & x\leq 0 \end{cases}$$

- This distribution is the distribution of the variance of random data

The Exponential Distribution

Definition

The *exponential distribution* is

$$f(x;\beta) = f(x;\alpha=1;\beta) = \begin{cases} \dfrac{1}{\beta}e^{-\frac{x}{\beta}} & x>0 \\ 0 & x\leq 0 \end{cases}$$

Where the random variable X is the time between events, given a mean time between events of $\beta = \dfrac{1}{r}$ where r is the mean rate of events

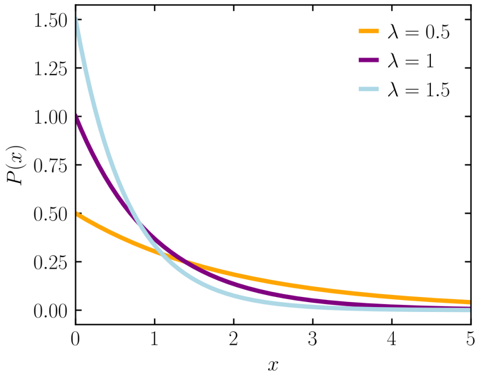

Figure 2: The exponential distribution ($\lambda = \beta$)

- Statistics:
  - $\mu = \beta$
  - $\sigma^2 = \beta^2$
- This is a decaying exponential which decays faster with larger β
  - Smaller values of β start off higher but decay more quickly
- This is the distribution of how long we need to wait for an event, given a mean waiting time of β
  - $\beta = \dfrac{1}{r}$ where r is the rate of events
  - Similar to a discrete version of the inverse binomial distribution
- Relation to the Poisson distribution: $p(x;\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$
  - The Poisson distribution gives us the distribution for the number of events in an interval of length t where $\lambda = rt$

  - The exponential distribution gives us the time between events
  - The probability of no events occurring in t is $p(0; rt) = e^{-rt}$, which is also the probability that the first event occurs after time t
  - Let X be the random variable for time to the first event, then $P(X \geq x) = p(0; rx) = e^{-rx}$
  - From this we can get a CDF $F(x) = P(X \leq x) = 1 - P(X \geq x) = 1 - e^{-rx}$
  - The PDF is then $\frac{\mathrm{d}}{\mathrm{d}x}F(x) = re^{-rx} = f\left(x; \beta = \frac{1}{r}\right)$
- Example: on an average day a component fails every $\beta = 4$ days; if the failures are described by an exponential distribution, what is the chance a component lasts more than a week?
  - We want $P(X > 7) = 1 - P(X \leq 7)$
  - This is given by $\int_7^\infty \frac{1}{\beta}e^{-\frac{t}{\beta}}\,\mathrm{d}t = 0 - (-e^{-\frac{7}{4}}) = 0.17$
  - Note the exponential distribution is a continuous probability distribution so we used an integral
- Related question: how many failures should we expect in a week?
  - This is a Poisson distribution with $\lambda = rt = \frac{1}{\beta}t = \frac{1}{4} \cdot 7 = \frac{7}{4}$
  - Therefore the expected number of failures is just $\frac{7}{4}$

Lecture 16, Feb 15, 2023

Memoryless Property of the Exponential Distribution

- Suppose we have a random variable X with exponential distribution $f(x) = \frac{1}{\beta}e^{-\frac{x}{\beta}}$
- Consider $P(X \geq s + t | X \geq s)$, i.e. if there has been no event for time s, what is the probability that there are no more events for another time t?
  - $$\begin{aligned} P(X \geq s + t | X \geq s) &= \frac{P(X \geq s + t \cap X \geq s)}{P(X \geq s)} \\ &= \frac{P(X \geq s + t)}{P(X \geq s)} \\ &= \frac{\int_{s+t}^\infty \frac{1}{\beta}e^{-\frac{x}{\beta}}\,\mathrm{d}x}{\int_s^\infty \frac{1}{\beta}e^{-\frac{x}{\beta}}\,\mathrm{d}x} \\ &= \frac{e^{-\frac{(s+t)}{\beta}}}{e^{-\frac{s}{\beta}}} \\ &= e^{-\frac{t}{\beta}} \\ &= P(X \geq t) \end{aligned}$$
- This means it doesn't matter how long we've waited – the past has no impact on the probability distribution in the future
  - This is known as the *memoryless* property
- Note that this is a modelling assumption that we have to be aware of; it is not a statement about reality

Functions of Random Variables

- Suppose X is a discrete random variable with PMF $f(x)$; suppose $Y = u(x)$ where u is one-to-one (aka bijective, invertible); what is the PMF of Y?
  - $X = u^{-1}(Y)$
  - $g(y) = P(Y = y)$ since the RVs are discrete
  - $g(y) = P(u^{-1}(Y) = u^{-1}(y)) = P(X = u^{-1}(y)) = f(u^{-1}(y))$
- If X is a discrete random variable with PDF $f(x)$, let $g(y)$ and $G(y)$ be the PDF and CDF of Y
  - We can no longer do the same thing as in the discrete case because $P(Y = y) = 0$, so we must consider the CDF

  - $G(y) = P(Y \leq y) = P(u^{-1}(Y) \leq u^{-1}(y)) = P(X \leq u^{-1}(y)) = \int_{-\infty}^{u^{-1}(y)} f(t)\,\mathrm{d}t$
  - To get back the PDF we need to differentiate
  - $g(y) = \dfrac{\mathrm{d}G}{\mathrm{d}y} = \dfrac{\mathrm{d}}{\mathrm{d}y}\int_{-\infty}^{u^{-1}(y)} f(t)\,\mathrm{d}t = f(u^{-1}(y))\dfrac{\mathrm{d}u^{-1}(y)}{\mathrm{d}y}$ using Leibniz's integral rule
  - Note here we made the assumption that $u(y)$ is strictly increasing; if it's strictly decreasing, then we need to flip the inequality in $G(y)$
    - * This means we need to add an absolute value around $\dfrac{\mathrm{d}u^{-1}}{\mathrm{d}y}$

Summary

Given a discrete random variable X with PMF $f(x)$, if $Y = u(x)$ where u is invertible, then the PMF of Y is given by

$$g(y) = f(u^{-1}(y))$$

If X is continuous, then we have

$$g(y) = f(u^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}u^{-1}(y)\right|$$

Functions of Multiple Random Variables and Non-Invertible Functions (Textbook Ch 7)

Theorem

Let X_1, X_2 be two discrete random variables with joint distribution $f(x_1, x_2)$; let

$$Y_1 = u_1(X_1, X_2), Y_2 = u_2(X_1, X_2)$$

define a one-to-one transformation, such that we may find unique inverse functions

$$x_1 = w_1(y_1, y_2), x_2 = w_2(y_1, y_2)$$

; then the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = f(w_1(y_1, y_2), w_2(y_1, y_2))$$

- If we just want the distribution of some $Y_1 = u_1(X_1, X_2)$, then we can make up a Y_2 such that we have a one-to-one transformation; this allows us to get $g(y_1, y_2)$, from which we can find $g(y_2) = \sum_{y_2} g(y_1, y_2)$
  - e.g. if $Y_1 = X_1 + X_2$ then we can let $Y_2 = X_2$, then our inverse functions are given by $x_1 = y_1 - y_2, x_2 = y_2$

Theorem

Let X_1, X_2 be two continuous random variables with joint distribution $f(x_1, x_2)$; let $Y_1 = u_1(X_1, X_2), Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation (like in the discrete case) with $x_1 = w_1(y_1, y_2), x_2 = w_2(y_1, y_2)$, then the joint probability distribution of Y_1 and Y_2 is given by

$$g(y_1, y_2) = f(w_1(y_1, y_2), w_2(y_1, y_2))|J|$$

where J is the Jacobian given by

$$J = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

- This theorem is a generalization of the one variable function of a continuous RV; in the case of a single RV, the Jacobian becomes simply the derivative of the inverse function u^{-1}

Theorem

Let X be a continuous random variable with distribution $f(x)$; let $Y = u(X)$ define a transformation such that u is not one-to-one, but the interval over which X is defined can be partitioned into k mutually disjoint sets such that each of the inverse functions

$$x_1 = w_1(y), x_2 = w_2(y), \cdots, x_k = w_k(y)$$

are one-to-one, then the probability distribution of Y is given by

$$g(y) = \sum_{i=1}^{k} f(w_i(y)) \left| \frac{\mathrm{d}w_i}{\mathrm{d}y} \right|$$

- Example: Let $Y = X^2$, then the inverse functions are $w_1 = -\sqrt{y}, w_2 = \sqrt{y}$ that divide the full range where X is defined
  - This gives us the derivatives $-\frac{1}{2\sqrt{y}}, \frac{1}{2\sqrt{y}}$
  - Therefore $g(y) = f(-\sqrt{y})\left|-\frac{1}{2\sqrt{y}}\right| + f(\sqrt{y})\left|\frac{1}{2\sqrt{y}}\right|$

Lecture 17 (Online)

Moments and Moment-Generating Functions

Definition

The rth *moment* about the origin of the random variable X is

$$\mu_r' = E[X^r] = \begin{cases} \sum_x x^r f(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} x^r f(x)\,\mathrm{d}x & X \text{ continuous} \end{cases}$$

- The first moment is the mean: $\mu = \mu_1'$
- The second moment is related to variance: $\sigma^2 = E[X^2] - \mu^2 = \mu_2' - \mu^2$

Definition

The *moment-generating function* of the random variable X is

$$M_X(t) = E[e^{tX}] = \begin{cases} \displaystyle\sum_x e^{tx} f(x) & X \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f(x) \, \mathrm{d}x & X \text{ continuous} \end{cases}$$

- Consider the discrete case: $\left.\dfrac{\mathrm{d}^r M_X(t)}{\mathrm{d}t^r}\right|_{t=0} = \left.\dfrac{\mathrm{d}^r}{\mathrm{d}t^r} \displaystyle\sum_x e^{tx} f(x)\right|_{t=0}$

$$\begin{aligned} &= \sum_x f(x) \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r}\right|_{t=0} \\ &= \sum_x f(x) x^r \left. e^{tx}\right|_{t=0} \\ &= \sum_x f(x) x^r \\ &= E[X^r] \\ &= \mu_r' \end{aligned}$$

  - This works the same in the continuous case
- In general $\mu_r' = \left.\dfrac{\mathrm{d}^r M_X(t)}{\mathrm{d}t^r}\right|_{t=0}$

Linear Combinations of Random Variables

- Consider a discrete RV X with distribution $f(x)$; let $Y = aX$, then the distribution $h(y) = f\left(\dfrac{y}{a}\right)$, using the formula we found before
- In the continuous case using the formula before $h(y) = \dfrac{1}{|a|} f\left(\dfrac{y}{a}\right)$
- If we have the moment generating function of X as $M_X(t)$, how do we find $M_Y(t)$?
  - $$\begin{aligned} M_Y(t) = \int_{-\infty}^{\infty} e^{ty} h(y) \, \mathrm{d}y &= \frac{1}{|a|} \int_{-\infty}^{\infty} e^{ty} f\left(\frac{y}{a}\right) \, \mathrm{d}y \\ &= \frac{1}{|\alpha|} \int_{-\infty}^{\infty} e^{taz} f(z) a \, \mathrm{d}z \\ &= \int_{-\infty}^{\infty} e^{taz} f(z) \, \mathrm{d}z \\ &= M_X(at) \end{aligned}$$
  - This is also true in the discrete case
- In general $M_{aX} = M_X(at)$
- What about a sum of independent RVs $Z = X + Y$?
  - $h(z) = P(X + Y = z) = \displaystyle\sum_w P(X = w) P(Y = z - w) = \sum_{w=-\infty}^{\infty} f(w) g(z - w)$
  - In the continuous case this is similar: $h(z) = \displaystyle\int_{-\infty}^{\infty} f(w) g(z - w) \, \mathrm{d}w = (f * g)(z)$
    - This is a convolution
  - $M_Z(t) = \displaystyle\sum_z e^{tz} h(z) = \sum_z e^{tz} \sum_w f(w) g(z - w) = \sum_w f(w) \sum_z e^{tz} g(z - w)$
    - Let $k = z - w$
    - $M_Z(t) = \displaystyle\sum_w f(w) \sum_k e^{t(k+w)} g(k) = \sum_w e^{tw} f(w) \sum_k e^{tk} g(k) = M_X(t) + M_Y(t)$
- In general $M_{X+Y} = M_X(t) M_Y(t)$

- There is a connection between moment generating functions and Laplace/Fourier transforms

Lecture 18, Feb 27, 2023

Application: Renewable Energy & Electricity Markets

- Renewable energy operators are paid for power produced
- Penalties for uncertainty in power production
- Basic quantities:
  - At the start of the hour, wind forecasts power $\hat{p}$
  - Actually produces power p
  - Market price λ
  - Fees u^+ for overproduction, u^- for underproduction
  - Total revenue is then $\tilde{J} = \lambda p - u^-(\hat{p} - p)^+ - u^+(p - \hat{p})^+$ where $(\cdot)^+ = \max(\cdot, 0)$
- Suppose p is distributed according to $f(p)$ and CDF $F(P)$; we want to maximize $J = E_p[\tilde{J}]$ over $\hat{p}$
- $E_p[\tilde{J}] = \lambda \int_{-\infty}^{\infty} pf(p)\,\mathrm{d}p - u^- \int_{-\infty}^{\hat{p}} (\hat{p} - p)f(p)\,\mathrm{d}p - u^+ \int_{\hat{p}}^{\infty} (p - \hat{p})f(p)\,\mathrm{d}p$
- $\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}\hat{p}} &= -u^-\left(\int_{-\infty}^{\hat{p}} f(p)\,\mathrm{d}p + (\hat{p} - \hat{p})f(p)\right) - u^+\left(\int_{\hat{p}}^{\infty} -f(p)\,\mathrm{d}p - (\hat{p} - \hat{p})f(p)\right) \\ &= -u^-F(\hat{p}) + u^+(1 - F(\hat{p}))\end{aligned}$
- Solve for $\hat{p}$: $-u^-F(\hat{p}) + u^+(1 - F(\hat{p})) = 0 \implies F(\hat{p}) = \dfrac{u^+}{u^- + u^+} \implies \hat{p} = F^{-1}\left(\dfrac{u^+}{u^- + u^+}\right)$
  - We know F is invertible, because by definition F is non-decreasing
- What does this tell us?
  - F caps at 1, so if $u^+ \gg u^-$, the ratio is close to 1, therefore $\hat{p} \to F^{-1}(1) = \infty$
    - If the penalty for overproducing is way bigger than the penalty for underproducing, then it's better to forecast a very big number
  - Conversely $u^+ \ll u^- \implies \hat{p} \to F^{-1}(0) = -\infty$
    - If the penalty for underproducing is way bigger than the penalty for overproducing, then it's better to forecast a very small number
  - If $u^- = u^+$ then $\hat{p} = F^{-1}\left(\dfrac{1}{2}\right)$ which is the *median* of the distribution of p
- This is classically known as the newsvendor problem

Lecture 19, Mar 1, 2023

Sampling

- Often we can't measure the entire population, so we instead examine a subset
- How representative is this subset to the population?
- A sample could be thought of as actual data $x_1, \cdots, x_n$, with the assumption that each x_i is a realization of an independent random variable X_i
  - Since each of these measurements is done separately, they are separate random variables, but we assume that they all have the same distribution as the population

Definition

The *sample mean* (or realized/empirical sample mean) is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The *random variable* of the mean is defined as

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

- $\bar{X}$ itself is a random variable
- Assuming all X_i are independent identically distributed (IID) with mean μ, then we can find $E[\bar{X}]$
  - $$\begin{aligned} E[\bar{X}] &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

Definition

The *sample variance* is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The *sample variance random variable* is defined as

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

- Why $n-1$ in the denominator?
  - The mean itself $\bar{x}$ is a function of the data points, so it is not independent; normally it would not have an effect on the variance, but in the case of sample variance, it has a contribution, which we eliminate via the $n-1$ in the denominator
  - Assume $\text{var}(X_i) = \sigma^2$ for all i (that is, each RV has the same variance); we would like to get $E[S^2] = \sigma^2$
  - $$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X} + \bar{X}^2 \\ &= \frac{1}{n-1}\left(n\bar{X}^2 - 2n\bar{X}^2 + \sum_{i=1}^{n} X_i^2\right) \\ &= \frac{1}{n-1}\left(-n\bar{X}^2 + \sum_{i=1}^{n} X^2\right) \end{aligned}$$

  - $$\begin{aligned} E[S^2] &= \frac{1}{n-1}\left(-nE[\bar{X}^2] + E\left[\sum_{i=1}^{n} X_i^2\right]\right) \\ &= \frac{1}{n-1}\left(-nE[\bar{X}^2] + \sum_{i=1}^{n} E[X_i^2]\right) \\ &= \frac{1}{n-1}\left(-n\left(E[\bar{X}]^2 + \operatorname{var}(\bar{X})\right) + \sum_{i=1}^{n} \mu^2 + \sigma^2\right) \\ &= \frac{1}{n-1}\left((-n\mu^2 + \operatorname{var}(\bar{X})) + n(\mu^2 + \sigma^2)\right) \\ &= \frac{1}{n-1}\left(-n\operatorname{var}(\bar{X}) + n\sigma^2\right) \\ &= \frac{1}{n-1}\left(-n\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + n\sigma^2\right) \\ &= \frac{1}{n-1}\left(-n\frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var}(X_i) + n\sigma^2\right) \\ &= \frac{1}{n-1}\left(-\frac{1}{n}n\sigma^2 + n\sigma^2\right) \\ &= \frac{1}{n-1}\left((n-1)\sigma^2\right) \\ &= \sigma^2 \end{aligned}$$
- We call this an *unbiased estimator* of the variance

Histograms

- In the case of a discrete RV, the x axis is the possible outcomes, the y axis is the number of times each outcome is observed
  - As the number of samples increases, the histogram divided by the sample size approaches the PMF
- In the case of a continuous RV we make bins to contain ranges of observations
  - As the number of samples increases and the bin size approaches 0, the histogram divided by the sample size approaches the PDF

Lecture 20, Mar 3, 2023

Random Sampling – Definitions

Definition

Population: The set of all possible observations, where each observation is a realization of a random variable

Sample: a subset of the population

- e.g. if we're measuring the heights of everybody in the world, then the population would be all the heights of everyone
- Each observation is the realization of a random variable

Definition

A *random sample* with n observations, where each observation is a realization of the random varaible $X_1, \cdots, X_n$, and we assume that $f(x_1, \cdots, x_n) = f(x_1) \cdots f(x_n)$; that is, the random variables are **independently and identically distributed**

A *statistic* is a function of the random variables $X_1, \cdots, X_n$

Definition

A sample is *biased* if it always leads to under or over-estimating some statistic of interest

- The distribution of a statistic is called a *sampling distribution* (e.g. the sampling distribution of $\bar{X}$ is $f(\bar{x})$)

Properties of Normal Random Variables

- Suppose X_1, X_2 are independent and normally distributed with means μ_1, μ_2, variances σ_1^2, σ_2^2, then $X_1 + X_2$ is still normally distributed, with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma = \sigma_1^2 + \sigma_2^2$
  - The mean and variance holds for any independent variable, but only when both X_1, X_2 are normal is $X_1 + X_2$ also normal
- $\frac{1}{n}X_1$ is also normal, with $\mu = \frac{\mu_1}{n}, \sigma^2 = \frac{\sigma_1^2}{n^2}$
- Suppose $X_1, \cdots, X_n$ are IID with μ, σ^2, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normal with mean μ and variance $\frac{\sigma^2}{n}$
  - This follows directly from the above 2 properties
  - Notice σ^2 decreases with increasing n – that is, the more data we have in our sample, the closer we will get to the true mean
  - The standard deviation decreases on the order of $\sqrt{n}$

The Central Limit Theorem

Theorem

Central Limit Theorem: Assume a sample with $X_1, \cdots, X_n$ identically and independently distributed with mean μ and finite variance σ^2 (with no restrictions on the distribution otherwise), where the sample mean is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, let

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

then as $n \to \infty$, the distribution of Z_n converges to the standard normal

$$\lim_{n \to \infty} f(z_n) = n(z_n; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_n^2}{2}}$$

That is, the distribution of $\bar{X}_n$ itself approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

- Z_n is never truly normally distributed but approaches a normal distribution as n gets bigger
- It doesn't matter what the actual distribution is, as we take more samples, the distribution of the average is going to look more and more like the Gaussian with smaller and smaller variance as the sample size increases

- The standard deviation of $\bar{X}_n$ is approximately $\dfrac{\sigma}{\sqrt{n}}$
- Example: a runner averages $\mu = 4$ minutes per mile with standard deviation $\sigma = 5$ seconds, what is the chance that the mean time of the next 20 races less than 3:58?
  - We want $P(\bar{X} < 238) = P\left(\dfrac{\bar{X} - 240}{\frac{5}{\sqrt{20}}} < \dfrac{-2}{\frac{5}{\sqrt{20}}}\right) = P(Z_{20} \leq -1.8) \approx \Phi(-1.8) = 0.036$

Lecture 21, Mar 6, 2023

Distribution of Sample Variance

- What is the distribution of the sample variance $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$?

Theorem

Let

$$\begin{aligned}\chi^2 &= \frac{(n-1)S^2}{\sigma^2} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

then χ^2 has a chi-squared distribution with $v = n - 1$, which is given by

$$f(y; v) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} y^{\frac{v}{2}-1}e^{-\frac{y}{2}} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

If μ is known, then

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

has a chi-squared distribution with $v = n$

- v is the number of degrees of freedom, or independent pieces of information
- In the case where $\bar{X}$ is used, because $\bar{X}$ itself is dependent on X_i, there is one fewer degree of freedom, which gives higher variance (chi-squared distribution shifts to the right)

The t-distribution

- Using CLT we can make inferences about the mean when σ^2 is known; however the t-distribution must be used when σ^2 is not known
- Consider the statistic $T = \dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$; for large n ($n \geq 30$) we have $S \approx \sigma$ so T approaches a normal distribution
- For a smaller n the t-distribution is a more accurate description

Definition

The t-distribution is given by

$$h(t;v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

Given samples $X_1, \cdots, X_n$ with sample mean $\bar{X}$ and sample variance S^2, then the statistic

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

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

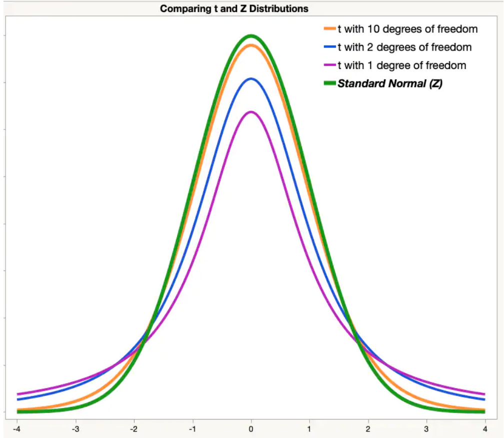

Figure 3: Shape of the t-distribution compared to the standard normal distribution

- The t-distribution has heavier "tails" than the standard normal – because we have less information, it's more likely that our estimate $\bar{X}$ is further from the true mean μ
- As the number of degrees of freedom $v \to \infty$ the t-distribution approaches the standard normal – if we have infinite samples, we'd know σ precisely

Lecture 22, Mar 8, 2023

CLT vs. t-distribution

- For the t-distribution, we need IID samples that are *normally distributed*; we don't know σ, so we use S instead, then $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ for *any* n
  - If $n \geq 30$ then $S \approx \sigma$ so CLT can be used even if the samples are non-normal
- For the CLT, we don't need to make any assumptions about the underlying distribution, but we need to know σ and have n big enough (if n is big enough we can approximate σ by S)
- The t-distribution is therefore less powerful because we need to make assumptions about the underlying distribution
- For the χ^2 distribution, we also need to assume a normal population
- Both χ^2 and t distributions are exact results, because we assume a normal underlying population; the CLT needs no such assumption, but it is an approximation

Quantile Plots

Definition

Let a sample $X_1, \cdots, X_n$, then a *quantile* $q(f)$ is defined such that a fraction f of the population is less than or equal to $q(f)$

In a *quantile plot*, f is plotted against $q(f)$

- e.g. $q(0.9) = 6'$ means that 90% of the population is less than 6' tall
- Example: data $-2, 0, 0, 1, 3, 3, 3, 4, 6$
  - Trick: plot $\left(\dfrac{i - \frac{3}{8}}{n + \frac{1}{4}}, x_i\right)$ for all i
- Observations:
  - $q(0.5)$ is the empirical sample *median*
  - $q(0.25)$ is the lower quartile, $q(0.75)$ is the upper quartile
  - Flat areas of a quantile plot indicates clusters of data that have the same value
- A quantile is the inverse of the cumulative distribution (however quantiles are not continuous and do not have to be strictly increasing)
  - Suppose we have X be an RV with CDF $F(x)$; $F(x)$ is $P(X \leq x)$ which is approximately the fraction of data less than or equal to x
  - The quantile $x = q(f)$ looks for x such that fraction f of data is less than or equal to x
  - Therefore $F(q(f)) = f$
  - If the quantile were a continuous and strictly increasing function of F, then $q = F^{-1}$
- A quantile distribution can be used to determine whether the data is normal
  - The quantile function for a normal distribution is $q(f) = \Phi^{-1}(f)$
    - * *Quantile function* always refers to the inverse of the standard normal CDF
  - This plot is going to look like the transpose of a standard normal CDF
  - Plot $x_1, \cdots, x_n$ on the vertical axis and $\Phi^{-1}(f_i) = q(f_i)$ where $f_i = \dfrac{i - \frac{3}{8}}{n + \frac{1}{4}}, i = 1, \cdots n$; if this plot is roughly linear, then the data is roughly normally distributed
    - * If we had a continuum of data, then we should have $q(f) = x$, so we would be plotting x_i against itself

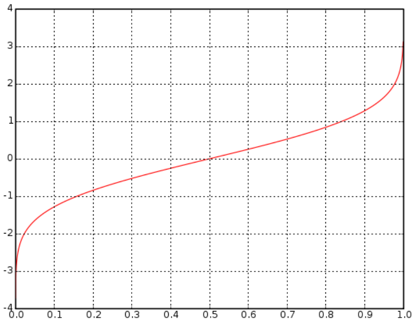

Figure 4: Plot of the quantile function

Lecture 23, Mar 9, 2023

Point Estimates

- So far we've worked with point estimates in our sampling
- We have IID measurements $X_1, \cdots, X_n$ with realizations $x_1, \cdots, x_n$
- In general we write θ is the true parameter, $\hat{\theta}$ is the observed value, and $\hat{\Theta}$ is the statistic
  - e.g. $\theta = \mu$ is the true parameter (true mean), $\hat{\theta} = \bar{x}$ is the observed value (observed mean), $\hat{\Theta} = \bar{X}$ is the statistic (sample mean RV)
- In general we want to estimate θ from $\hat{\theta}$

Definition

$\hat{\Theta}$ is an *unbiased estimator* if $E[\hat{\Theta}] = \theta$, that is, the expectation of the statistic is the true mean

Out of all unbiased estimators, the most *efficient* estimator has the lowest variance

- e.g. with the mean, $E[X_i] = \mu$ for all i, so any of the individual estimates is an unbiased estimator; however $\bar{X}$ has lower variance (σ^2/n vs. σ^2), so $\bar{X}$ is the most efficient estimator of the sample mean

Interval Estimates

- Instead of estimating an exact value, interval estimates give an interval $\theta_L \leq \theta \leq \theta_U$
  - The most well known example are confidence intervals
  - This gives us a sense of how good our estimate is
- θ_L, θ_U should be the realization of some sampling statistic based on the data

Definition

A *confidence interval* is of the form

$$P(\Theta_L \leq \theta \leq \Theta_U) = 1 - \alpha$$

where Θ_L, Θ_U are statistics

- e.g. a 95% confidence interval has $\alpha = 0.05$
- To calculate confidence intervals of the mean we can use the CLT
  - $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
  - By the CLT the distribution of Z approaches $n(z; 0, 1)$
  - Recall the CDF is $\displaystyle\int_{-\infty}^{x} n(z; 0, 1)\,\mathrm{d}z$
  - Let $\beta < 0.5$, define z_β such that $\Phi(-z_\beta) = \beta \implies z_\beta = -\Phi^{-1}(\beta)$, that is, the area under the normal PDF above $x = \beta$ is equal to α
    - By symmetry $1 - \Phi(z_\beta) = \beta$
  - $1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$
    - $z_{\alpha/2}$ has α area above it, and α area below $-z_{\alpha/2}$ by our previous definition
  - $1 - \alpha = P\left(-z_{\alpha/2} \leq \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$
  - This gives us $\Theta_L = \bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \Theta_U = \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

Definition

Given data $x_1, \cdots, x_n$, let

$$\theta_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\theta_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$, such that $\alpha/2$ of the normal distribution is below $-z_{\alpha/2}$, then $[\theta_L, \theta_U]$ is the *observed* (or realized) confidence interval for the mean for some confidence $1 - \alpha$

- Example: $n = 20, \bar{x} = 4$, we know $\sigma = 2$, we want a 95% confidence interval
  - $\alpha = 0.05 \implies z_{\alpha/2} = -\Phi^{-1}(0.025) = 1.96 \approx 2$
  - $0.95 = P\left(\bar{X} - 2\frac{2}{\sqrt{20}} \leq \mu \leq \bar{X} - 2\frac{2}{\sqrt{20}}\right) = P(\bar{X} - 0.88 \leq \mu \leq \bar{X} + 0.88)$
  - Therefore our realized confidence interval is $[4 - 0.88, 4 + 0.88] = [3.12, 4.88]$
- This does **not** mean that there is a 95% chance that the true mean falls within $[3.12, 4.88]$ (this statement is not mathematically valid since the true mean is not a random variable)
  - It means that if we did this experiment a large number of times, each time collecting 20 samples, 95% of the time the realization $[\bar{X} - 0.88, \bar{X} + 0.88]$ contains the true mean
  - When a confidence interval is reported as a pair of numbers, as $[3.12, 4.88]$, it is only a particular realization of the confidence interval for that particular experiment

Lecture 24, Mar 13, 2023

Other Confidence Intervals

- We may want a one-sided confidence interval $1 - \alpha = P(Z \leq z_\alpha)$
  - Same approach may be used to get $1 - \alpha = P\left(\mu \leq \bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right) = P(\mu \leq \bar{X}_U)$
- If the variance is unknown we have to use the t-distribution, thereby assuming the population is normal
  - Let $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ with sample variance $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
    - T has a t-distribution, $h(t)$
    - Let the CDF be $H(t) = \int_{-\infty}^{t} h(x)\,\mathrm{d}x$
  - Once again define $t_\beta > 0$ for $\beta < 0.5$ such that $H(-t_\beta) = \beta$, that is, the area under the PDF above t_β is β
  - $1 - \alpha = P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$
    - Therefore we find $\bar{X}_L = \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}_U = \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$
  - Note if we had known variance we end up getting a smaller confidence interval

Standard Error

- Notice in our confidence interval we used $\frac{\sigma}{\sqrt{n}}$; the size of the interval is proportional to it
- $\frac{\sigma}{\sqrt{n}}$ is referred to as the *standard error*, since in a way it tells us how much error there could be in our estimate of the mean

Prediction Intervals

- If we have samples $X_1, \cdots X_n$, normally distributed, what can we say about the next single sample X_0?
- $\bar{X}$ is a good point estimate, so what is the error $X_0 - \bar{X}$?
  - We know the distribution of this error is normal, with variance $\sigma^2 + \dfrac{\sigma^2}{n}$
- Let $Z = \dfrac{\bar{X}_0 - \bar{X}}{\sigma\sqrt{1 + \frac{1}{n}}}$, then Z has distribution $n(z; 0, 1)$
- Therefore $1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \implies P\left(\bar{X} - z_{\alpha/2}\sigma\sqrt{1 + \dfrac{1}{n}} \leq X_0 \leq \bar{X} + z_{\alpha/2}\sigma\sqrt{1 + \dfrac{1}{n}}\right)$
- This gives us our *prediction interval*

Definition

Given samples $X_1, \cdots, X_n$ normally distributed, the *prediction interval* is defined as

$$\left[\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}, \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}\right]$$

such that

$$1 - \alpha = P\left(\bar{X} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \leq X_0 \leq \bar{X} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}\right)$$

That is, there is a probability $1 - \alpha$ that the next single sample X_0 falls within this interval

- Prediction intervals can be useful for outlier detection

Lecture 25, Mar 15, 2023

Tolerance Intervals

Definition

Given n identically and independently distributed observations $X_1, \cdots, X_n$, the tolerance interval is defined as

$$[\bar{x} - ks, \bar{x} + ks]$$

with k chosen such that a fraction $1 - \alpha$ of the population is within the interval; that is,

$$\lim_{n \to \infty} P(-kS \leq X \leq kS) = 1 - \alpha$$

where X is an observation of the population, $\bar{X}$ is the sample mean and S^2 is the sample variance

- The tolerance interval n doesn't shrink with n
- Values of k can be obtained from tables

Summary

3 types of intervals:

1. Confidence intervals: $1-\alpha$ chance of the true mean μ being in this interval around $\bar{x}$; with increasing n, this shrinks to 0, because $\bar{x}$ approaches the true mean
2. Prediction intervals: $1-\alpha$ chance of the next observation x_0 being in this interval around $\bar{x}$; with increasing n, the interval shrinks to a fixed value $\bar{x} \pm z_{\alpha/2}\sigma$
3. Tolerance limits: For large n, fraction $1-\alpha$ of all measurements will be in this interval around $\bar{x}$; k does not change with n, but as n increases the relationship becomes more precise

Two Samples, Known Variance

- Consider 2 samples with sizes n_1, n_2, each having mean μ_1, μ_2 and known variances σ_1^2, σ_2^2
- $\bar{X}_1 - \bar{X}_2$ is normal with mean $\mu_1 - \mu_2$, variance $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$ since $\bar{X}_1, \bar{X}_2$ are normal by the CLT
- Then $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ follows the standard normal by CLT

Two Samples, Unknown Variance

- First consider the case where the variances are equal but unknown
  - If $n_1, n_2 > 30$ then we can use s_1, s_2 as estimates of σ_1, σ_2 and use the CLT as usual
  - If $n_1, n_2 < 30$, we have to use the t-distribution and assume normal population
  - Use variance $S_p^2 = \dfrac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
    - This pooled variance estimate is a sample size-weighted average of the two individual sample variances
  - Let $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
  - T has a t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom
- If $\sigma_1 \neq \sigma_2$ and both are unknown
  - Let $T' = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
  - T' approximately has a t-distribution
  - $v = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$
  - If this is not an integer, round it down to the nearest integer (Satterthwaite approximation)

Lecture 26, Mar 16, 2023

Monte Carlo Integration

- What if we want to numerically integrate a function in a large number of dimensions? $\displaystyle\int_{x_1} \cdots \int_{x_d} f(x_1, \cdots, x_d)\,\mathrm{d}x_d \cdots \mathrm{d}x_1$
  - In 1 dimension we can use the trapezoidal rule
  - The number of points we need to take scales exponentially with the number of dimensions, so for higher dimensions the number of points we need becomes impractical
  - The error of the trapezoidal approximation scales as $O(N^{-2})$ where N is the number of points per dimension; given N total points and d dimensions, the error would be $O((N^{1/d})^{-2}) = O(N^{-2/d})$

- As the number of dimensions increase, the computational effort increases exponentially – "the curse of dimensionality"
- Consider if we had a sample of N uniform random variables in $[0,1]^d$, i.e. a d-dimensional unit cube, each random variable being a d-dimensional vector
  - This is a sample $\begin{bmatrix} x_1^1 \\ \vdots \\ x_d^1 \end{bmatrix}, \begin{bmatrix} x_1^2 \\ \vdots \\ x_d^2 \end{bmatrix}, \cdots, \begin{bmatrix} x_1^N \\ \vdots \\ x_d^N \end{bmatrix}$
- Let RV $Y = f(X_1, \cdots, X_d)$, where f is the function we want to integrate
  - $E[Y] = \int_0^1 \cdots \int_0^1 f(x_1, \cdots, x_d)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d$
    - Since the distribution is uniform the distribution is 1 for all points in the unit cube
    - The expectation turns out to be the same as the integral we wanted to evaluate
- Let $\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y^k = \frac{1}{N}\sum_{k=1}^{N} f(x_1^k, \cdots, x_d^k)$; this is our sample mean
  - $E[\bar{Y}] = \frac{1}{N}\sum_{k=1}^{N} E[f(x_1^k, \cdots, x_d^k)] = \frac{1}{N}\sum_{k=1}^{N} \mu = \mu$
    - This means $\bar{Y}$ is an unbiased estimator of $E[Y]$, which has the same value of the integral we want to solve
- Therefore we can take the average of a bunch of uniform samples in the unit cube to estimate the integral (referred to as *Monte Carlo* integration)
- Let $Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{N}}}$, then by the CLT as $N \to \infty$, Z has standard normal distribution
  - The standard deviation of $\bar{Y} - \mu$ is decreasing at a rate of $\frac{1}{\sqrt{n}}$
  - This means the error is also decreasing at a rate of $\frac{1}{\sqrt{N}} = N^{-\frac{1}{2}}$
- The full steps of Monte Carlo integration:
  1. Make uniform random variables $\begin{bmatrix} x_1^k \\ \vdots \\ x_d^k \end{bmatrix}, k = 1, \cdots, N$ (the sample points)
  2. Evaluate the function at these points $Y^k = f(x_1^k, \cdots, x_d^k)$
  3. The result of the integral is approximately $\frac{1}{N}\sum_{k=1}^{N} Y^k$
- Compare MC's order of $O(N^{-\frac{1}{2}})$ to trapezoidal rule's $O(N^{-\frac{2}{d}})$
  - MC's rate of convergence is unaffected by dimension while trapezoidal rule is highly dependent on dimensionality
  - For lower number of dimensions ($d < 4$) trapezoidal rule is still better
  - For any higher dimensions MC becomes better

Lecture 27, Mar 20, 2023

Paired Observations

- *Paired observations* are when we have 2 populations and 2 samples of the same size (1 from each sample); in this case we can take one sample from each population and pair them up
  - e.g. measuring a chemical reactor, taking measurements at the inlet and outlet at the same time; a medical experiment where we measure before and after for each person
- Let $(X_i, Y_i), i = 1, \cdots, n$ be the paired samples; we're interested in the difference $D_i = X_i - Y_i$
  - $\mathrm{var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$
  - If we assume X and Y to be negatively correlated, we expect $\sigma_{XY} > 0$; this reduces the total

variance

  * Compared to treating X and Y as independent and taking the difference of means, this gives lower variance
  * Due to lower variance this gives tighter confidence intervals

Note

The gain in quality of the confidence interval of pairing vs. not pairing will be the greatest when there is homogeneity within units (strong correlation between two observations in a pair) and large differences between different units.

Pairing effectively reduces the number of degrees of freedom, so it may actually be counterproductive if the reduction in variance is small.

Confidence Intervals for Binomial Distributions

- Consider n IID trials, with $P(Y_i = 1) = p, P(Y_i = 0) = 1 - p$ for $i = 1, \cdots, n$; $X = \sum_{i=1}^{n} Y_i$ is the number of 1s, giving the binomial PMF $b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$, $\mu = np, \sigma^2 = np(1-p)$
- Let $Z = \frac{X - np}{\sqrt{np(1-p)}}$ then by CLT as $n \to \infty$, the PDF of Z becomes $n(x; 0, 1)$
  - This is because X is the sum of a series of Bernoulli RVs so the CLT applies
- Can we estimate $p = P(Y_i = 1)$?
  - Use $\hat{P} = \frac{X}{n}$ as the estimator
  - $\mu_{\hat{P}} = E\left[\frac{X}{n}\right] = \frac{np}{n} = p$ so the estimator is unbiased
  - $\sigma^2_{\hat{P}} = \operatorname{var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\operatorname{var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$
  - Let $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ and by CLT this approaches the standard normal
  - Confidence interval is more challenging because it's harder to isolate for p
- $1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$

$$= P\left(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right)$$

  - We have 2 choices: solve quadratics for p to get an exact confidence interval, or if p is large, approximate $p = \hat{p} = \frac{x}{n}$
  - $1 - \alpha = P\left(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}\right)$

$$= P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

  - Here $\hat{p} = \frac{x}{n}$ is not a random variable, but $\hat{P}$ is
  - This approximation relies on p not being too close to 0 or 1; as a heuristic, both $n\hat{p}$ and $n\hat{q}$ should be at least 5
- How big does n need to be to have $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \delta$?
  - Solving for n, we get $n > \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{\delta^2}$, but $\hat{p} = \frac{x}{n}$

        - * If we can get a crude estimate of p, we can use that to first determine n
        - * n should be rounded up
    - $\hat{p}(1-\hat{p})$ is bounded by $1/4$ since $\hat{p} \leq 1$
    - We can have a safe lower bound by $n \geq \dfrac{z_{\alpha/2}^2}{4\delta^2}$

Lecture 28, Mar 22, 2023

Lecture 29, Mar 23, 2023

Confidence Interval of the Variance

- Recall that $W^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$ follows a chi-squared distribution with $v = n-1$ degrees of freedom, assuming a normal distribution
- The chi-squared distribution is asymmetric, so getting a confidence interval is harder
- Let χ_β^2 be the value of χ^2 such that the area under the distribution to the left of it is β
    - $\chi_{\alpha/2}^2$ is the value of χ^2 such that the area under the distribution to the *left* of it is $\alpha/2$
    - $\chi_{1-\alpha/2}$ is the value of χ^2 such that the area under the distribution to the *right* of it is $\alpha/2$
- Denote the CDF of the chi-squared distribution as $F(y; v) = \int_0^y f(x; v)\,\mathrm{d}x$
    - $\chi_{\alpha/2}^2 = F^{-1}(\alpha/2; v)$ and $\chi_{1-\alpha/2}^2 = F^{-1}(1-\alpha/2; v)$
- $1 - \alpha = P(\chi_{\alpha/2}^2 \leq W^2 \leq \chi_{1-\alpha/2}^2)$

$$
\begin{aligned}
&= P\left(\chi_{\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2\right) \\
&= P\left(\frac{\chi_{\alpha/2}^2}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi_{1-\alpha/2}^2}{(n-1)S^2}\right) \\
&= P\left(\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2}^2}\right)
\end{aligned}
$$

Equation

Confidence interval of the variance: Given n IID samples, with a sample variance of S^2 and a confidence level of $1-\alpha$, then the confidence interval for the true variance σ^2 is

$$\left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{\alpha/2}^2}\right]$$

where $\chi_\beta^2 = F^{-1}(\beta; v)$, and F is the CDF of the chi-squared distribution with $v = n-1$ degrees of freedom

Maximum Likelihood Estimation

- So far we've relied on intuition to define our estimators (e.g. $\bar{X}$ for μ, S^2 for σ^2, etc)
- Can we find a systematic way to define an estimator for any statistic? (e.g. what if we wanted to estimate v in a chi-squared distribution?)

Definition

The *likelihood function* for an IID sample $x_1, \cdots, x_n$, with each sample distributed according to a PDF $g(x;\theta)$, where θ is a parameter vector, is

$$\begin{aligned} L(x_1, \cdots, x_n; \theta) &= f(x_1, \cdots, x_n; \theta) \\ &= g(x_1;\theta) \cdots g(x_n;\theta) \end{aligned}$$

The *maximum likelihood estimator* is then

$$\hat{\theta} = \max_{\theta} L(x_1, \cdots, x_n; \theta)$$

- Maximum likelihood estimation estimates the parameter by attempting to maximize the likelihood function, which roughly describes the probability of getting the particular sample
  - In the discrete case, $f(x_1, \cdots, x_n; \theta)$ is exactly the probability of the sample occurring; with a continuous distribution it is more complicated but the intuition still holds
- Example: $n = 1, x_1 = 3, \theta = \mu$, standard normal $f(x_1; \theta) = n(x_1; \theta, 1)$
  - We're trying to move the mean around so that $f(x_1; \theta)$ is maximized
  - The optimal value is $\theta = x_1 = 3$ because the normal distribution peaks at its mean
  - i.e. having a mean of 3 makes it the most likely that we'll get $x_1 = 3$
- Example: Bernoulli distribution, estimating p, given sample $1, 0, 1, 1$
  - We can make this a binomial distribution with 3 successes
  - $L(1, 0, 1, 1; p) = \binom{4}{3} p^3 (1-p)^1 \propto p^3(1-p) = p^3 - p^4$
  - $\dfrac{\mathrm{d}L}{\mathrm{d}p} \propto 3p^2 - 4p^3 = 0 \implies 3 - 4p = 0 \implies p = \dfrac{3}{4}$
  - This is exactly what we expect – if we get 3 successes in 4 trials, then we'd estimate the success probability to be $\dfrac{3}{4}$

Lecture 30, Mar 27, 2023

Log-Likelihood

- With distributions such as the normal, exponential, or Poisson distributions, we can make them easier to work with if we instead maximize the log of the likelihood function, since $\dfrac{\mathrm{d}}{\mathrm{d}x} \ln x > 0$
- Example: normal distribution, find μ, σ^2
  - $\begin{aligned} L(x_1, \cdots, x_n) &= \prod_{i=1}^{n} n(x_i; \mu, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2} \end{aligned}$
  - $\begin{aligned} \ln L &= \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\ &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 \end{aligned}$

- $$\frac{\partial L}{\partial \mu} = 0$$
  $$\implies \frac{1}{2}\sum_{i=1}^{n} 2\left(\frac{x_i - \mu}{\sigma}\right)^2 = 0$$
  $$\implies \sum_{i=1}^{n} x_i - \mu = 0$$
  $$\implies \left(\sum_{i=1}^{n} x_i\right) - n\mu = 0$$
  $$\implies \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
  - We just ended up getting $\bar{x}$, which makes sense
- $$\frac{\partial L}{\partial \sigma^2} = 0$$
  $$\implies -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$
  $$\implies \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$
  - This is an unbiased estimator if we do know the true mean μ, but if we use the optimal $\mu = \bar{x}$, this would be a biased estimator
  - In the limit as $n \to \infty$ the sample mean approaches the true mean, and the estimator becomes unbiased

> **Important**
>
> When we use maximum likelihood estimation for each parameter, it is assumed that we know the exact values of the other parameters; this means that in the case of multiple parameters, MLE can give a biased estimate

> **Note**
>
> There is often a tradeoff between bias and variance of an estimator. Sometimes it might be worth it to have an estimator that is a little biased if it significantly reduces the variance of the estimator, making it more efficient.

Lecture 31, Mar 29, 2023

Hypothesis Testing

- A *hypothesis* is a conjecture made about a population
  - e.g. $H_0 : P(H) = 0.5$ for a coin; an alternative hypothesis is $H_1 : P(H) > 0.5$
- In general, we have a *null hypothesis* H_0 (the status quo, or what we believe before the experiment), and an *alternative hypothesis* H_1; with each sample tested, we either reject H_0 for H_1, or we fail to reject H_0 and nothing changes
  - In the case where we fail to reject H_0, we can't make conclusions because it's still possible that H_0 is rejected in a later experiment
  - e.g. in a drug trial, H_0 is the drug has no effect, H_1 is the drug having an effect

Type I and II Errors

- *Type I* errors are rejections of H_0 when it is true (i.e. false positives); α is the probability of a type 1 error
  - These result from oversensitive tests
- *Type II* errors are failures to reject H_0 when it is false (i.e. false negatives); β is the probability of a type 2 error
  - These results from undersensitive tests
- Example: testing for mean tensile strength of a new alloy
  - $H_0 : \mu = 1000\text{MPa}$
  - $H_1 : \mu \neq 1000\text{MPa}$
  - Suppose we have $n = 25$ with $\sigma = 50$, to use the CLT
  - First, $P(\mu = 1000) = 0$ since μ is continuous; we therefore need to define a range where we don't reject H_0, e.g. $990 \leq \bar{x} \leq 1010$
    * The *critical region* is the complement of this range (i.e. the range that results in rejection of H_0)
  - This is a confidence interval; we want to compute the confidence
  - $\alpha = P(\text{Type I})$

    $$\begin{aligned} &= P(\bar{X} < 990 \cup \bar{X} > 1010 \mid \mu = 1000) \\ &= 1 - P(990 \leq \bar{X} \leq 1010 \mid \mu = 1000) \\ &= 1 - P\left(\frac{990 - 1000}{\frac{50}{\sqrt{25}}} \leq Z \leq \frac{1010 - 1000}{\frac{50}{\sqrt{25}}}\right) \\ &= 1 - P(-1 \leq Z \leq 1) \\ &= 0.32 \end{aligned}$$

    * The chance of a false positive is 32%, which is not good
    * To reduce α, we can either increase n, or widen the range where H_0 is accepted
    * If we do the latter however, that makes our test less sensitive and increases the probability of a type II error
- The exact size of the critical region is often ad-hoc, but consistency is key – do many tests with the same critical region, and compare results

Lecture 32, Mar 31, 2023

Hypothesis Testing Continued

- Continuing the example from the last lecture:
  - To find β, i.e. the chance of concluding $\mu \in [990, 1010]$ when this is not the case, we first have to assume some value of μ outside this interval; this becomes somewhat arbitrary
    * Unlike in the case of α where we have a single set value of μ, now μ can be in a range and the choice of μ will affect the result
  - Assume $\mu = 1020$ then $z_U = \dfrac{1010 - 1020}{10} = -1, z_L = \dfrac{990 - 1020}{10} = -3$
  - $\beta = P(-3 \leq Z \leq -1) = \Phi(-1) - \Phi(-3) \approx 0.15$
  - To decrease β, we can again increase n, move our assumption of μ, or shrinking the interval
  - There is a fundamental tradeoff between α and β – changing the interval will decrease one but increase the other
  - Again, the choice of assumption of μ is very ad-hoc, but the important thing is consistency; do many tests with the same critical region and assumption of μ, then the results can be compared

Lecture 33, Apr 3, 2023

One vs. Two-Tailed Tests

- *Two-tailed tests* are for hypotheses for an exact value (or two-sided range) of a variable; *one-tailed tests* are for hypotheses with one-sided ranges
- So far we've only considered two-tailed tests (e.g. temperature tomorrow is 10°C; H_0 is temperature is 10 degrees, H_1 is temperature is not)
- A one-tailed test would be e.g. hypothesis is eating something would cause you to live past 80, then $H_0 : \theta \leq 80, H_1 : \theta > 80$, we can set critical region $\theta > 82$

Relation of Hypothesis Testing to Confidence Intervals

- Consider a scenario where $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$, for a sample of size n with known variance σ^2
- Let $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ as usual
- Before, we set the critical region and then calculated the probability of a type I error; what happens if we set α first and then calculate the critical region?
- $\alpha = P(\text{Type I}) = P(Z \leq -z_{\alpha/2} \cup Z \geq z_{\alpha/2}) \implies 1 - \alpha = P\left(-z_{\alpha/2} \leq \dfrac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right)$
- The critical region would be $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \notin [-z_{\alpha/2}, z_{\alpha/2}]$
- This is the exact same as a confidence interval

p-Values

Definition

Given a sample $X_1, \cdots, X_n$ with variance σ^2, and the null hypothesis $H_0 : \mu = \mu_0$, then the p-value is defined as

$$\begin{aligned} p &= P\left(\left|\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| > |z|\right) \\ &= P(|Z| \geq |z|) \\ &= 2P(Z > |z|) \\ &= 2(1 - \Phi(|z|)) \end{aligned}$$

where $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ has distribution $n(z; 0, 1)$ and z is the observed value of Z

- The p value is the probability that, given H_0 is true, an even more extreme observation will be made
  - If $p \approx 0$, that means z is very far from 0 and therefore $\bar{x}$ is very far from μ_0, so H_0 is highly unlikely
  - If $p \approx 1$, then z is close to 0 so $\bar{x} \approx \mu$, so H_0 cannot be rejected
- We still need to assume H_0, but unlike the other methods of hypothesis testing, to compute a p-value we don't need to make an ad-hoc critical region
- Example: $H_0 : \mu = 5, H_1 : \mu \neq 5$, with $n = 40, \bar{x} = 5.5, s = \sigma = 1$
  - $z = \dfrac{5.5 - 5}{\frac{1}{\sqrt{40}}} \approx 3.16$
    * If we choose $\alpha = 0.05$ so the critical region for Z is outside $[-1.96, 1.96]$, then z is outside the critical region so we reject H_0
  - $p = 2P(Z > |z|) = 2(1 - \Phi(3.16)) = 0.0016 \approx 0$
    * The very low p-value means H_0 is highly unlikely

  - Note to compute the p-value we didn't need to specify a critical region

Lecture 34, Apr 5, 2023

Linear Regression

- We have a set of data in the form of input-output pairs, $(x_i, y_i), i = 1, \dots, n$; in general we want a function f such that $y = f(x)$ minimizes the errors $e_i = y_i - f(x_i)$
- For now we will talk about linear regression – assuming $y = ax + b$ so $e_i = y_i - (ax_i - b)$, so the total squared error is $\mathcal{E} = \displaystyle\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - ax_i - b)^2$
- Goal: find $\displaystyle\min_{a,b} \mathcal{E}$
  - $$\begin{aligned}\frac{\partial \mathcal{E}}{\partial a} &= \sum_{i=1}^{n} \frac{\partial}{\partial a}(y_i - ax_i - b)^2 \\ &= -\sum_{i=1}^{n} 2(y_i - ax_i - b)x_i \\ &= 0\end{aligned}$$
  - $$\begin{aligned}\frac{\partial \mathcal{E}}{\partial b} &= \sum_{i=1}^{n} \frac{\partial}{\partial b}(y_i - ax_i - b)^2 \\ &= -\sum_{i=1}^{n} 2(y_i - ax_i - b) \\ &= 0\end{aligned}$$
  - Rearrange and we get the normal equations: $\begin{cases} nb + a\displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b\displaystyle\sum_{i=1}^{n} x_i + a\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$
  - Since they are linearly independent we can directly solve; let $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i, \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
  - Solve: $\begin{cases} a = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b = \bar{y} - a\bar{x} \end{cases}$
  - We know this is a minimum, since $\mathcal{E}$ is a simple convex quadratic
- With normal linear regression we only look at the vertical distances (i.e. errors in y); but we can improve it by looking at the normal (geometric) distance instead, which is called a *Deming regression*
  - In order to do this we also need to know the ratio of variances

Least Squares With Maximum Likelihood Estimation

- We assume each error e_i is a realization of a normal RV with mean 0 and variance σ^2
- $$\begin{aligned}L(e_1, \dots, e_n; a, b) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}} \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}}\end{aligned}$$
- If we maximize this, we get the same solution as the least squares approach
- We can think of this as assuming that each y_i is normally distributed, with mean $\mu = ax_i - b$ and uniform variance σ^2

Lecture 35, Apr 10, 2023

Support Vector Machines

- In normal regression we had $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$; for classification problems, they generally have input $x \in \mathbb{R}^n$ with output $y \in \{-1, 1\}$, i.e. the output is a binary yes or no
- We can express a hyperplane as $\boldsymbol{w}^T\boldsymbol{x} - b = 0$, where $\boldsymbol{w}$ is the normal vector defining the orientation and b is an offset from the origin
- The hyperplane divides all of the input space into 2 regions, $\boldsymbol{w}^T\boldsymbol{x} > b$ and $\boldsymbol{w}^T\boldsymbol{x} < b$; each region corresponds to a different value of y
- Given some data, we're looking for a hyperplane that separates the 2 types of data
  - We also want a hyperplane that's the most "in the middle" and divides the empty space between 2 types evenly
- We want to find $\boldsymbol{w}^T\boldsymbol{x} - b = 0$ that maximizes d, the distance on each side of the hyperplane, while separating the data
- Unlike in linear regression, this problem is not analytically solvable
  - In linear regression, a change in any data point is going to affect the total error and therefore change the solution; however in this problem moving a data point may not affect the solution at all
  - Depending on the orientation of the hyperplane $\boldsymbol{w}$, different data points will become relevant
- What is the expression for d?
  - Consider 2 parallel hyperplanes, $\boldsymbol{w}^T\boldsymbol{x} = 1$ and $\boldsymbol{w}^T\boldsymbol{x} = 0$ and some point on the first hyperplane so $\boldsymbol{w}^T\boldsymbol{x}^* = 1$, then $d = \|\boldsymbol{x}^*\|$
  - We know $\boldsymbol{x}^* = \alpha\boldsymbol{w}$ so $\|\boldsymbol{x}^*\| = \dfrac{1}{\|\boldsymbol{w}\|}$, $\alpha = \dfrac{1}{\|\boldsymbol{w}\|^2} \implies \boldsymbol{x}^* = \dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|^2}$
  - $d = \dfrac{1}{\|\boldsymbol{w}\|}$
- To maximize d we want to minimize $\|\boldsymbol{w}\|$ or $\|\boldsymbol{w}\|^2$, subject to the constraint that if $y_i = 1$, then $\boldsymbol{w}^T\boldsymbol{x}_i - b > 0$, or if $y_i = -1$, then $\boldsymbol{w}^T\boldsymbol{x}_i - b < 0$
- The support vector machine is $\min\limits_{\boldsymbol{w},b}\|\boldsymbol{w}\|^2$ such that $y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 0$ for all training data
  - This is a quadratic program
  - If this is solvable, i.e. the data is separable, then we have an optimal classifier

Lecture 36, Apr 12, 2023

Markov Chains

- Let us have n states, and $p_i(k)$ which gives the probability of being in state i in time k, for $i = 1, \dots, n$; at any time if we're in state i, then there is a probability P_{ij} of moving to state j
  - Therefore $\displaystyle\sum_{i=1}^{n} p_i(k) = 1$ since the system must be in some state at any time
  - $\displaystyle\sum_{j=1}^{n} P_{ij} = 1$ since the system must go somewhere in the next time (this includes itself, i.e. $P_{ii} \geq 0$)
- Observe that $p_i(k+1) = \displaystyle\sum_{j=1}^{n} p_j(k)P_{ji}$, i.e. the probability of being in i at the next time is the sum of the probabilities of all states transitioning to i
- Let $\boldsymbol{p}(k) = \begin{bmatrix} p_1(k) \\ \vdots \\ p_n(k) \end{bmatrix}, \boldsymbol{M} = \begin{bmatrix} P_{00} & \cdots & P_{n0} \\ \vdots & \ddots & \vdots \\ P_{0n} & \cdots & P_{n0} \end{bmatrix}$, then $\boldsymbol{p}(k+1) = \boldsymbol{M}\boldsymbol{p}(k)$, analogous to an LTI system
  - Generally $\boldsymbol{p}(k+s) = \boldsymbol{M}^s\boldsymbol{p}(k)$
  - Notice that $\mathbf{1}^T\boldsymbol{M} = \mathbf{1}^T$ where $\mathbf{1}$ is the vector of all 1s, so it's a left eigenvector with eigenvalue 1
- Any eigenvector $\boldsymbol{q} = \boldsymbol{M}\boldsymbol{q}$ is a steady state, which the system will never come out of once entered

  - We can show that $\boldsymbol{q} = \lim_{k\to\infty} \boldsymbol{M}^k\boldsymbol{p}$ for any initial PMF $\boldsymbol{p}$, if such $\boldsymbol{q}$ exists
  - Intuitively if we just let the Markov chain do its thing eventually it'll end up in a steady state
- Example: Suppose we have 2 states, on (1) or off (0), $P_{11} = 0.99, P_{10} = 0.01, P_{01} = 0, P_{00} = 1$
  - $\boldsymbol{M} = \begin{bmatrix} 1 & 0.01 \\ 0 & 0.99 \end{bmatrix}$
  - Steady state is $\boldsymbol{p} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, i.e. 100% chance of off