Statistics for Data Science - 2

Week 1 Important formulas

Basic Probability

1. **Experiment:** Process or phenomenon that we wish to study statistically.
   Example: Tossing a fair coin.

2. **Outcome:** Result of the experiment.
   Example: head is an outcome on tossing a fair coin.

3. **Sample space:** A sample space is a set that contains all outcomes of an experiment.
   • Sample space is a set, typically denoted S of an experiment.
   • example: Toss a coin: S = { heads, tails }

4. **Event:** An event is a subset of the sample space.

   - Toss a coin: S = { heads, tails }
     - Events: empty set, {heads}, {tails}, { heads, tails }
     - 4 events
   - An event is said to have "occurred" if the actual outcome of the experiment belongs to the event.
   - One event can be contained in another, i.e. $A \subseteq B$
   - Complement of an event A, denoted A^C= { outcomes in S not in A } = $(S \setminus A)$.
   - Since events are subsets, one can do complements, unions, intersections.

5. **Disjoint events:** Two events with an empty intersection are said to be disjoint events.

   - Throw a die: even number, odd number are disjoint.
   - Multiple events: $E_1, E_2, E_3, ...$ are disjoint if, for any $i \neq j$, $E_i \cap E_j$ = empty set.

6. **De Morgan's laws:** For any two events A and B,
   $(A \cup B)^C = A^C \cap B^C$ and $(A \cap B)^C = A^C \cup B^C$.

7. **Probability:** "Probability" is a unction P that assigns to each event a real number between 0 and 1 and satisfies the following two axioms:

   (i) $P(S) = 1$ (probability of the entire sample space equals 1).

   (ii) If $E_1, E_2, E_3, ...$ are disjoint events (Could be infinitely many),

   $$P(E_1 \cup E_2 \cup E_3 \cup ...) = P(E_1) + P(E_2) + P(E_3) + ...$$

   - Probability function Assigns a value that represents chance of occurrence of the event.

- Higher value of the probability of an event means higher chance of occurring that event.
- 0 means event cannot occur and 1 means event always occurs.

8. Probability of the empty set (denoted ϕ) equals 0. that is

$$P(\phi) = 0$$

9. Let E^C be the complement of Event E. Then,

$$P(E^C) = 1 - P(E)$$

10. If event E is the subset of event F, that is $E \subseteq F$, then

$$P(F) = P(E) + P(F \setminus E)$$

$$\Rightarrow P(E) \leq P(F)$$

11. If E and F are events, then

$$P(E) = P(E \cap F) + P(E \setminus F)$$

$$P(F) = P(E \cap F) + P(F \setminus E)$$

12. If E and F are events, then

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

13. **Equally likely events:** assign the same probability to each outcome.

14. If sample space S contains the equally likely outcomes, then
    - $P(\text{one outcome}) = \dfrac{1}{\text{Number of outcomes in } S}$
    - $P(\text{event}) = \dfrac{\text{Number of outcomes in event}}{\text{Number of outcomes in } S}$

15. **Conditional probability space:** Consider a probability space (S, E, P), where S represents the sample space, E represents the collection of events, and P represents the probability function.

    - Let B be an event in S with $P(B) > 0$. Now, conditional probability space given B is defined as
      For any event A in the original probability space (P, S, E), the conditional probability of A given B is $\dfrac{P(A \cap B)}{P(B)}$.
    - It is denoted by $P(A \mid B)$. And
    $$P(A \cap B) = P(B)P(A \mid B)$$

16. **Law of total probability:**

    - If the events B and B^c partitioned the sample space S such that $P(B_1), P(B_2) \neq 0$, then for any event A of S,
    $$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$
    - In general, if we have k events $B_1, B_2, \cdots, B_k$ that partition S, then for any event A in S,
    $$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(A \mid B_i)P(B_i).$$

17. **Bayes' theorem:** Let A and B are two events such that $P(A) > 0, P(B) > 0$.
    $$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$
    $$\Rightarrow P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)}$$
    In general, if the events $B_1, B_2, \cdots, B_k$ partition S such that $P(B_i) \neq 0$ for $i = 1, 2, \cdots, k$, then for any event A in S such that $P(A) \neq 0$,
    $$P(B_r \mid A) = \frac{P(B_r)P(A \mid B_r)}{\sum\limits_{i=1}^{k} P(B_i)P(A \mid B_i)}$$
    for $r = 1, 2, \cdots, k$.

18. **Independence of two events:** Two events A and B are independent iff
    $$P(A \cap B) = P(A)P(B)$$
    - A and B independent $\Rightarrow P(A \mid B) = P(A)$ and $(B \mid A) = P(B)$ for $P(A), P(B) > 0$.

- Disjoint events are never independent.
- A and B independent $\Rightarrow$ A and B^c are independent.
- A and B independent $\Rightarrow$ A^c and B^c are independent.

19. **Mutual independence of three events:** Events A, B, and C are mutually independent if
    (a) $P(A \cap B) = P(A)P(B)$
    (b) $P(A \cap C) = P(A)P(C)$
    (c) $P(A \cap B) = P(A)P(B)$
    (d) $P(A \cap B \cap C) = P(A)P(B)P(C)$

20. **Mutual independence of multiple events:** Events $A_1, A_2, \cdots, _n$ are mutually independent if, $\forall i_1, i_2, \cdots, i_k$,
$$P(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k} \cap) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$
    n events are mutually independent $\Rightarrow$ any subset with or without complementing are independent as well.

21. Occurrence of event A in a sample space is considered as *success.*

22. Non - occurrence of event A in a sample space is considered as *failure.*

23. **Repeated independent trials:**
    (a) **Bernoulli trials**
    - Single Bernoulli trial:
      - Sample space is {success, failure} with P(success) = p.
      - We can also write the sample space S as $\{0, 1\}$, where 0 denotes the failure and 1 denotes the success with $P(1) = p, P(0) = 1 - p$.
        This kind of distribution is denoted by $Bernoulli(p)$.
    - Repeated Bernoulli trials:
      - Repeat a Bernoulli trial multiple times independently.
      - For each of the trial, the outcome will be either 0 or 1.

    (b) **Binomial distribution:** Perform n independent $Bernoulli(p)$ trials.
    - It models the number of success in n independent Bernoulli trials.
    - Denoted by $B(n, p)$.
    - Sample space is $\{0, 1, \cdots, n\}$.
    - Probability distribution is given by
$$P(B(n, p) = k) = nC_k p^k (1 - p)^{n-k}$$
      where n represents the total number trials and k represent the number of success in n trials.

- $P(B=0) + P(B=1) + \cdots + P(B=n) = 1$
  $\Rightarrow (1-p)^n + nC_2p^2(1-p)^{n-2} + \cdots + p^n = 1.$

(c) **Geometric distribution:** It models the number of failures the first success.

- Outcomes: Number of trials needed for first success and is denoted by $G(p)$.
- Sample space: $\{1, 2, 3, 4, \cdots\}$
- $P(G = k) = P(\text{first } k-1 \text{ trials result in 0 and } kth \text{ trial result in 1.}) = (1-p)^{k-1}p.$
- Identity: $P(G \leq k) = 1 - (1-p)^k$.

Statistics for Data Science - 2

Week 2 Important formulas

1. **Random variable:** A random variable is a function with domain as the sample space of an experiment and range as the real numbers, i.e. a function from the sample space to the real line.
   - Toss a coin, Sample space = $\{H, T\}$
     - Random variable $X : X(H) = 0, X(T) = 1$
2. **Random variables and events:** If X is a random variable,
   $(X < x) = \{s \in S : X(s) < x\}$ is an event for all real x.
   So, $(X > x), (X = x), (X \leq x), (X \geq x)$ are all events.
   - Throw a die, Sample space = $\{1, 2, 3, 4, 5, 6\}$
     - $E = 0$: event $\{1, 3, 5\}$
     - $E = 1$: event $\{2, 4, 6\}$
     - $E < 0$: null event
     - $E \leq 1$: event $\{1, 2, 3, 4, 5, 6\}$
3. **Range of a random variable:** The range of a random variable is the set of values taken by it. Range is a subset of the real line.
   - Throw a die, $E = 0$ if number is odd, $E = 1$ if number is even
     - Range = $\{0, 1\}$
4. **Discrete random variable:** A random variable is said to be discrete if its range is a discrete set.
5. **Probability Mass Function (PMF):** The probability mass function (PMF) of a discrete random variable (r.v.) X with range set T is the function $f_X : T \to [0, 1]$ defined as
   $f_X(t) = P(X = t)$ for $t \in T$.
6. **Properties of PMF:**
   - $0 \leq f_X(t) \leq 1$
   - $\sum_{t \in T} f_X(t) = 1$
7. **Uniform random variable:** $X \sim$ Uniform(T), where T is some finite set.

- Range: Finite set T
- PMF: $f_X(t) = \frac{1}{|T|}$ for all $t \in T$

8. **Bernoulli random variable:** $X \sim \text{Bernoulli}(p)$, where $0 \leq p \leq 1$.
   - Range: $\{0, 1\}$
   - PMF: $f_X(0) = 1 - p, f_X(1) = p$

9. **Binomial random variable:** $X \sim \text{Binomial}(n, p)$, where n: positive integer, $0 \leq p \leq 1$.
   - Range: $\{0, 1, 2, \dots, n\}$
   - PMF: $f_X(k) = {}^nC_k p^k (1-p)^{n-k}$

10. **Geometric random variable:** $X \sim \text{Geometric}(p)$, where $0 < p \leq 1$.
    - Range: $\{1, 2, \dots, n\}$
    - PMF: $f_X(k) = (1-p)^{k-1} p$

11. **Negative Binomial random variable:** $X \sim \text{Negative Binomial}(r, p)$, where r: positive integer, $0 < p \leq 1$.
    - Range: $\{r, r+1, r+2, \dots\}$
    - PMF: $f_X(k) = {}^{k-1}C_{r-1} (1-p)^{k-r} p^r$

12. **Poisson random variable:** $X \sim \text{Poisson}(\lambda)$, where $\lambda > 0$.
    - Range: $\{0, 1, 2, 3, \dots\}$
    - PMF: $f_X(k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$

13. **Hypergeometric random variable:** $X \sim \text{HyperGeo}(N, r, m)$, where N, r, m: positive integers
    - Range: $\{\max(0, m-(N-r)), \dots, \min(r, m)\}$
    - PMF: $f_X(k) = \dfrac{{}^rC_k \, {}^{N-r}C_{m-k}}{{}^NC_m}$

14. **Functions of a random variable:** X : random variable with PMF $f_X(t)$.
    $f(X)$: random variable whose PMF is given as follows.

$$\begin{aligned} f_{f(X)}(a) &= P(f(X) = a) = P(X \in \{t : f(t) = a\}) \\ &= \sum_{t: f(t) = a} f_X(t) \end{aligned}$$

   - PMF of $f(X)$ can be found using PMF of X.

Statistics for Data Science - 2

Week 3 Notes

Multiple Random Variables

1. **Joint probability mass function:** Suppose X and Y are discrete random variables defined in the same probability space. Let the range of X and Y be T_X and T_Y , respectively. The joint PMF of X and Y , denoted f_{XY} , is a function from $T_X \times T_Y$ to $[0, 1]$ defined as

$$f_{XY}(t_1, t_2) = P(X = t_1 \text{ and } Y = t_2), t_1 \in T_X, t_2 \in T_Y$$

   - Joint PMF is usually written as table or a matrix.
   - $P(X = t_1 \text{ and } Y = t_2)$ is denoted $P(X = t_1, Y = t_2)$

2. **Marginal PMF:** Suppose X and Y are jointly distributed discrete random variables with joint PMF f_{XY}. The PMF of the individual random variables X and Y are called as marginal PMFs. It can be shown that

$$f_X(t_1) = P(X = t_1) = \sum_{t_2 \in T_Y} (f_{XY}(t_1, t_2))$$

$$f_Y(t_2) = P(X = t_2) = \sum_{t_1 \in T_X} (f_{XY}(t_1, t_2))$$

   Note: Given the joint PMF, the marginal is unique.

3. **Conditional distribution given an event:** Suppose X is a discrete random variable with range T_X , and A is an event in the same probability space. The conditional PMF of X given A is defined as the PMF

$$f_{X|A}(t) = P(X = t|A)$$

   where $t \in T_X$
   We will denote the conditional random variable by $X|A$. (Note that $X|A$ is a valid random variable with PMF $f_{X|A}$).

   - $f_{X|A}(t) = \dfrac{P((X = t) \cap A)}{P(A)}$
   - Range of $(X|A)$ can be different from T_X and will depend on A.

4. **Conditional distribution of one random variable given another:**
Suppose X and Y are jointly distributed discrete random variables with joint PMF f_{XY} . The conditional PMF of Y given $X = t$ is defined as the PMF

$$f_{Y|X=x}(y) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f_{XY}(x, y)}{f_X(x)}$$

We will denote the conditional random variable by $Y|(X = x)$. (Note that $Y|(X = x)$ is a valid random variable with PMF $f_{Y|(X=x)}$.

- Range of $(Y|X = t)$ can be different from T_Y and will depend on t.
- $f_{XY}(x, y) = f_{Y|X=x}(x, y).f_X(x) = f_{X|Y=y}(x, y).f_Y(y)$
- $\sum\limits_{y \in T_Y} f_{Y|X=x}(y) = 1$

5. **Joint PMF of more than two discrete random variables:**
Suppose $X_1, X_2, \ldots, X_n$ are discrete random variables defined in the same probability space. Let the range of X_i be T_{X_i}. The joint PMF of X_i, denoted by $f_{X_1 X_2 \ldots X_n}$, is a function from $T_{X_1} \times T_{X_2} \times \ldots \times T_{X_n}$ to [0, 1] defined as

$$f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n) = P(X_1 = t_1, X_2 = t_2, \ldots, X_n = t_n); t_i \in T_{X_i}$$

6. **Marginal PMF in case of more than two discrete random variables:**
Suppose $X_1, X_2, \ldots, X_n$ are jointly distributed discrete random variables with joint PMF $f_{X_1 X_2 \ldots X_n}$. The PMF of the individual random variables $X_1, X_2, \ldots, X_n$ are called as marginal PMFs. It can be shown that

$$f_{X_1}(t_1) = P(X_1 = t_1) = \sum_{t_2 \in T_{X_2}, t_3 \in T_{X_3}, \ldots, t_n \in T_{X_n}} f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n)$$

$$f_{X_2}(t_2) = P(X_2 = t_2) = \sum_{t_1 \in T_{X_1}, t_3 \in T_{X_3}, \ldots, t_n \in T_{X_n}} f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n)$$

$$\vdots$$

$$f_{X_n}(t_n) = P(X_n = t_n) = \sum_{t_1 \in T_{X_1}, t_2 \in T_{X_2}, \ldots, t_{n-1} \in T_{X_{n-1}}} f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n)$$

7. **Marginalisation:** Suppose $X_1, X_2, \ldots, X_n$ are jointly distributed discrete random variables with joint PMF $f_{X_1 X_2 \ldots X_n}$. The joint PMF of the random variables $X_{i_1}, X_{i_2}, \ldots X_{i_k}$, denoted by $f_{X_{i_1} X_{i_2} \ldots X_{i_k}}$ is given by

$$f_{X_{i_1} X_{i_2} \ldots X_{i_k}}(t_{i_1}, t_{i_2}, \ldots t_{i_k}) = \sum f_{X_1 X_2 \ldots X_n}(t_1, \ldots t_{i_1 - 1}, t_{i_1}, t_{i_1 + 1}, \ldots t_{i_k - 1}, t_{i_k}, t_{i_k + 1} \ldots, t_n)$$

- Sum over everything you don't want.

8. **Conditioning with multiple discrete random variables:**

   - A wide variety of conditioning is possible when there are many random variables. Some examples are:
   - Suppose $X_1, X_2, X_3, X_4 \sim f_{X_1X_2X_3X_4}$ and $x_i \in T_{X_i}$, then
     - $f_{X_1|X_2=x_2}(x_1) = \dfrac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}$
     - $f_{X_1,X_2|X_3=x_3}(x_1, x_2) = \dfrac{f_{X_1X_2X_3}(x_1, x_2, x_3)}{f_{X_3}(x_3)}$
     - $f_{X_1|X_2=x_2,X_3=x_3}(x_1) = \dfrac{f_{X_1X_2X_3}(x_1, x_2, x_3)}{f_{X_2X_3}(x_2, x_3)}$
     - $f_{X_1X_4|X_2=x_2,X_3=x_3}(x_1, x_4) = \dfrac{f_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4)}{f_{X_2X_3}(x_2, x_3)}$

9. **Conditioning and factors of the joint PMF:**
   Let $X_1, X_2, X_3, X_4 \sim f_{X_1X_2X_3X_4}, X_i \in T_{X_i}$.

$$\begin{aligned}
f_{X_1X_2X_3X_4}(t_1, t_2, t_3, t_4) =& P(X_1 = t_1 \text{ and } (X_2 = t_2, X_3 = t_3, X_4 = t_4)) \\
=& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1) P(X_2 = t_2 \text{ and } (X_3 = t_3, X_4 = t_4)) \\
=& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1) f_{X_2|X_3=t_3,X_4=t_4}(t_2) P(X_3 = t_3 \text{ and } X_4 = t_4) \\
=& f_{X_1|X_2=t_2,X_3=t_3,X_4=t_4}(t_1) f_{X_2|X_3=t_3,X_4=t_4}(t_2) f_{X_3|X_4=t_4}(t_3) f_{X_4}(t_4).
\end{aligned}$$

   - Factoring can be done in any sequence.

10. **Independence of two random variables:**
    Let X and Y be two random variables defined in a probability space with ranges T_X and T_Y , respectively. X and Y are said to be independent if any event defined using X alone is independent of any event defined using Y alone. Equivalently, if the joint PMF of X and Y is f_{XY}, X and Y are independent if

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

    for $x \in T_X$ and $y \in T_Y$

    - X and Y are independent if

$$f_{X|Y=y}(x) = f_X(x)$$
$$f_{Y|X=x}(y) = f_Y(y)$$

      for $x \in T_X$ and $y \in T_Y$

    - To show X and Y independent, verify

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

      for **all** $x \in T_X$ and $y \in T_Y$

- To show X and Y dependent, verify

$$f_{XY}(x,y) \neq f_X(x) f_Y(y)$$

for **some** $x \in T_X$ and $y \in T_Y$
   - **Special case:** $f_{XY}(t_1,t_2) = 0$ when $f_X(t_1) \neq 0, f_Y(t_2) \neq 0$.

11. **Independence of multiple random variables:**
Let $X_1, X_2, \ldots, X_n$ be random variables defined in a probability space with range of X_i denoted T_{X_i}. $X_1, X_2, \ldots, X_n$ are said to be independent if events defined using different X_i are mutually independent. Equivalently, $X_1, X_2, \ldots, X_n$ are independent iff

$$f_{X_1 X_2 \ldots X_n}(t_1, t_2, \ldots, t_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

for all $x_i \in T_{X_i}$

   - All subsets of independent random variables are independent.

12. **Independent and Identically Distributed (i.i.d.) random variables:**
Random variables $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributed (i.i.d.), if
(i) they are independent.
(ii) the marginal PMFs f_{X_i} are identical.
Examples:

   - Repeated trials of an experiment creates i.i.d. sequence of random variables
     - Toss a coin multiple times.
     - Throw a die multiple times.
   - Let $X_1, X_2, \ldots X_n \sim$ i.i.d.X (Geometric(p)).
X will take values in $\{1, 2, \ldots\}$
$P(X = k) = p^{k-1}p$

     Since X_i's are independent and identically distributed, we can write

$$\begin{aligned} P(X_1 > j, X_2 > j, \ldots, X_n > j) =& P(X_1 > j) P(X_2 > j) \ldots P(X_n > j) \\ =& [P(X > j)]^n \end{aligned}$$

$$\begin{aligned} P(X > j) =& \sum_{k=j+1}^{\infty} (1-p)^{k-1} p \\ =& (1-p)^j p + (1-p)^{j+1} p + (1-p)^{j+2} p + \ldots \\ =& (1-p)^j p [1 + (1-p) + (1-p)^2 + \ldots] \\ =& (1-p)^j p \left(\frac{1}{1-(1-p)} \right) \\ =& (1-p)^j \end{aligned}$$

$$\Rightarrow P(X_1 > j, X_2 > j, \ldots, X_n > j) = [P(X > j)]^n = (1-p)^{jn}$$

13. **Function of random variables ($g(X_1, X_2, \ldots, X_n)$):**
Suppose $X_1, X_2, \ldots, X_n$ have joint PMF $f_{X_1X_2\ldots X_n}$ with T_{X_i} denoting the range of X_i. Let $g : T_{X_1} \times T_{X_2} \times \ldots \times T_{X_n} \to R$ be a function with range T_g . The PMF of $X = g(X_1, X_2..., X_n)$ is given by

$$f_X(t) = P(g(X_1, X_2..., X_n) = t) = \sum_{(t_1,\ldots,t_n):g(X_1,X_2...,X_n)=t} f_{X_1X_2\ldots X_n}(t_1, t_2, \ldots, t_n)$$

- **Sum of two random variables taking integer values:**
$X, Y \sim f_{XY}, Z = X + Y$.
Let z be some integer,

$$\begin{aligned}
P(Z = z) =& P(X + Y = z) \\
&= \sum_{x=-\infty}^{\infty} P(X = x, Y = z - x) \\
&= \sum_{x=-\infty}^{\infty} f_{XY}(x, z - x) \\
&= \sum_{y=-\infty}^{\infty} f_{XY}(z - y, y)
\end{aligned}$$

- **Convolution:** If X and Y are independent, $f_{X+Y}(z) = \sum\limits_{x=-\infty}^{\infty} f_X(x)f_Y(z - x)$

- Let $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$
  - X and Y are independent.
  - $Z = X + Y$, $z \in \{0, 1, 2, \ldots\}$

  $f_Z(z) \sim \text{Poisson}(\lambda_1 + \lambda_2)$
  $(X = k \mid Z = n) \sim \text{Binomial}\left(n, \dfrac{\lambda_1}{\lambda_1 + \lambda_2}\right), (Y = k \mid Z = n) \sim \text{Binomial}\left(n, \dfrac{\lambda_2}{\lambda_1 + \lambda_2}\right)$

14. **CDF of a random variable:**
Cumulative distribution function of a random variable X is a function $F_X : R \to [0, 1]$ defined as

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

15. **Minimum of two random variables:**
Let $X, Y \sim f_{XY}$ and let $Z = \min\{X, Y\}$, then

-

$$\begin{aligned}
f_Z(z) = P(Z = z) &= P(\min\{X, Y\} = z) \\
&= P(X = z, Y = z) + P(X = z, Y > z) + P(X > z, Y = z) \\
&= f_{XY}(z, z) + \sum_{t_2>z} f_{XY}(z, t_2) + \sum_{t_1>z} f_{XY}(t_1, z)
\end{aligned}$$

- $$\begin{aligned}F_Z(z) = P(Z \leq z) &= P(\min\{X,Y\} \leq z)\\ &= 1 - P(\min\{X,Y\} > z)\\ &= 1 - [P(X > z, Y > z)]\end{aligned}$$

16. **Maximum of two random variables:**
    Let $X, Y \sim f_{XY}$ and let $Z = \max\{X,Y\}$, then

    - $$\begin{aligned}f_Z(z) = P(Z = z) &= P(\max\{X,Y\} = z)\\ &= P(X = z, Y = z) + P(X = z, Y < z) + P(X < z, Y = z)\\ &= f_{XY}(z,z) + \sum_{t_2 < z} f_{XY}(z, t_2) + \sum_{t_1 < z} f_{XY}(t_1, z)\end{aligned}$$

    - $$\begin{aligned}F_Z(z) = P(Z \leq z) &= P(\max\{X,Y\} \leq z)\\ &= [P(X \leq z, Y \leq z)]\end{aligned}$$

17. **Maximum and Minimum of n i.i.d. random variables**

    - Let $X \sim \text{Geometric}(p), Y \sim \text{Geometric}(q)$
      X and Y are independent.
      $Z = \min(X,Y)$
      $$Z \sim \text{Geometric}(1 - (1-p)(1-q))$$

    - Maximum of 2 **independent** geometric random variables is not geometric.

Important Points:

1. Let $N \sim \text{Poisson}(\lambda)$ and $X|N = n \sim \text{Binomial}(n,p)$, then $X \sim \text{Poisson}(\lambda p)$

2. Memory less property of Geometric(p)
   If $X \sim \text{Geometric}(p)$, then
   $$P(X > m + n | X > m) = P(X > n)$$

3. Sum of n **independent** Bernoulli(p) trials is Binomial(n,p).

4. Sum of 2 **independent** Uniform random variables is not Uniform.

5. Sum of **independent** Binomial(n,p) and Binomial(m,p) is Binomial($n+m,p$).

6. Sum of r **i.i.d.** Geometric(p) is Negative-Binomial(r, p).

7. Sum of **independent** Negative-Binomial(r, p) and Negative-Binomial(s, p) is Negative-Binomial($r + s, p$)

8. If X and Y are independent, then $g(X)$ and $h(Y)$ are also independent.

Statistics for Data Science - 2

Week 4 Notes

Expected value

- **Expected value of a random variable**
  Definition: Suppose X is a discrete random variable with range T_X and PMF f_X . The expected value of X, denoted $E[X]$, is defined as

$$E[X] = \sum_{t \in T_X} tP(X = t)$$

  assuming the above sum exists.
  Expected value represents "center" of a random variable.

  1. Consider a constant c as a random variable X with $P(X = c) = 1$.
  $$E[c] = c \times 1 = c$$
  2. If X takes only non-negative values, i.e. $P(X \geq 0) = 1$. Then,
  $$E[X] \geq 0$$

- **Expected value of a function of random variables**
  Suppose $X_1 \ldots X_n$ have joint PMF $f_{X_1 \ldots X_n}$ with range of X_i denoted as T_{X_i}. Let

$$g : T_{X_1} \times \ldots \times T_{X_n} \to \mathbb{R}$$

  be a function, and let $Y = g(X_1, \ldots, X_n)$ have range T_Y and PMF f_Y . Then,

$$E[g(X_1, \ldots, X_n)] = \sum_{t \in T_Y} t f_Y(t) = \sum_{t_i \in T_{X_i}} g(t_1, \ldots, t_n) f_{X_1 \ldots X_n}(t_1, \ldots, t_n)$$

- **Linearity of Expected value:**
  1. $E[cX] = cE[X]$ for a random variable X and a constant c.
  2. $E[X + Y] = E[X] + E[Y]$ for any two random variables X, Y.

- **Zero mean Random variable:**
  A random variable X with $E[X] = 0$ is said to be a zero-mean random variable.

- **Variance and Standard deviation:**
  Definition: The variance of a random variable X, denoted by $\text{Var}(X)$, is defined as

$$\text{Var}(X) = E[(X - E[X])^2]$$

Variance measures the spread about the expected value.
Variance of random variable X is also given by $\mathrm{Var}(X) = E[X^2] - E[X]^2$

The standard deviation of X, denoted by $SD(X)$, is defined as

$$SD(X) = +\sqrt{\mathrm{Var}(X)}$$

Units of $SD(X)$ are same as units of X.

- **Properties: Scaling and translation**
  Let X be a random variable. Let a be a constant real number.
  1. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$
  2. $SD(aX) = \mid a \mid SD(X)$
  3. $\mathrm{Var}(X + a) = \mathrm{Var}(X)$
  4. $SD(X + a) = SD(X)$

- **Sum and product of independent random variables**
  1. For any two random variables X and Y (independent or dependent), $E[X+Y] = E[X] + E[Y]$.
  2. If X and Y are independent random variables,
     (a) $E[XY] = E[X]E[Y]$
     (b) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

- **Standardised random variables:**
  1. Definition: A random variable X is said to be standardised if $E[X] = 0, \mathrm{Var}(X) = 1$.
  2. Let X be a random variable. Then, $Y = \dfrac{X - E[X]}{SD(X)}$ is a standardised random variable.

- **Covariance:**
  Definition: Suppose X and Y are random variables on the same probability space. The covariance of X and Y, denoted as $\mathrm{Cov}(X, Y)$, is defined as

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

  It summarizes the relationship between two random variables.
  Properties:
  1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
  2. $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$

3. Covariance is symmetric if $\text{Cov}(X, Y) = Cov(Y, X)$
4. Covariance is a "linear" quantity.
   (a) $\text{Cov}(X, aY + bZ) = a\text{Cov}(X, Y) + b\text{Cov}(X, Z)$
   (b) $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$
5. Independence: If X and Y are independent, then X and Y are uncorrelated, i.e. $\text{Cov}(X, Y) = 0$
6. If X and Y are uncorrelated, they may be dependent.

- **Correlation coefficient:**
  Definition: The correlation coefficient or correlation of two random variables X and Y , denoted by $\rho(X, Y)$, is defined as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)}$$

1. $-1 \leq \rho(X, Y) \leq 1$.
2. $\rho(X, Y)$ summarizes the trend between random variables.
3. $\rho(X, Y)$ is a dimensionless quantity.
4. If $\rho(X, Y)$ is close to zero, there is no clear linear trend between X and Y.
5. If $\rho(X, Y) = 1$ or $\rho(X, Y) = -1$, Y is a linear function of X.
6. If $\mid \rho(X, Y) \mid$ is close to one, X and Y are strongly correlated.

- **Bounds on probabilities using mean and variance**

1. Markov's inequality: Let X be a discrete random variable taking non-negative values with a finite mean μ. Then,

$$P(X \geq c) \leq \frac{\mu}{c}$$

   Mean μ, through Markov's inequality: bounds the probability that a non-negative random variable takes values much larger than the mean.

2. Chebyshev's inequality: Let X be a discrete random variable with a finite mean μ and a finite variance σ^2. Then,

$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

   Other forms:

   (a) $P(\mid X - \mu \mid \geq c) \leq \dfrac{\sigma^2}{c^2}, P((X - \mu)^2 > k^2\sigma^2) \leq \dfrac{1}{k^2}$
   (b) $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \dfrac{1}{k^2}$

   Mean μ and standard deviation σ, through Chebyshev's inequality: bound the probability that X is away from μ by $k\sigma$.

Statistics for Data Science - 2

Week 5 Notes

Continuous Random Variables

1. **Cumulative distribution function:**
   A function $F : \mathbb{R} \to [0, 1]$ is said to be a Cumulative Distribution Function (CDF) if
   (i) F is a non-decreasing function taking values between 0 and 1.
   (ii) As $x \to -\infty$, $F \to 0$
   (iii) As $x \to \infty$, $F \to 1$
   (iv) Technical: F is continuous from the right.

2. **CDF of a random variable:**
   Cumulative distribution function of a random variable X is a function $F_X : R \to [0, 1]$ defined as
   $$F_X(x) = P(X \leq x)$$
   Properties of CDF
   - $F_X(b) - F_X(a) = P(a < X \leq b)$
   - F_X is a non-decreasing function of x.
   - F_X takes non-negative values.
   - As $x \to -\infty$, $F_X(x) \to 0$
   - As $x \to \infty$, $F_X(x) \to 1$

3. **Theorem: Random variable with CDF F(x)**
   Given a valid CDF $F(x)$, there exists a random variable X taking values in $\mathbb{R}$ such that
   $$P(X \leq x) = F(x)$$
   - If F is not continuous at x and $F(X)$ rises from F_1 to F_2 at x (jump at x), then
   $$P(X = x) = F_2 - F_1$$
   - If F is continuous at x, then
   $$P(X = x) = 0$$

4. **Continuous random variable:**
   A random variable X with CDF $F_X(x)$ is said to be a continuous random variable if $F_X(x)$ is continuous at every x.
   Properties of continuous random variables
   - CDF has no jumps or steps.
   - $P(X = x) = 0$ for all x.

- Probability of X falling in an interval will be nonzero

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

- Since $P(X = a) = 0$ and $P(X = b) = 0$, we have

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

5. **Probability density function (PDF):**
   A continuous random variable X with CDF $F_X(x)$ is said to have a PDF $f_X(x)$ if, for all x_0,
   $$F_X(x_0) = \int_{-\infty}^{x_0} f_X(x)dx$$
   - CDF is the integral of the PDF.
   - Derivative of the CDF (wherever it exists) is usually taken as the PDF.
   - Value of PDF around $f_X(x_0)$ is related to X taking a value around x_0.
   - Higher the PDF, higher the chance that X lies there.

6. For a random variable X with PDF f_X, an event A is a subset of the real line and its probability is computed as
   $$P(A) = \int_A f_X(x)dx$$
   - $P(a < X < b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$

7. **Density function:**
   A function $f : \mathbb{R} \to \mathbb{R}$ is said to be a density function if
   (i) $f(x) \geq 0$
   (ii) $\int_{-\infty}^{\infty} f_X(x)dx = 1$
   (iii) $f(x)$ is piece-wise continuous

8. Given a density function f, there is a continuous random variable X with PDF as f.

9. **Support of random variable X**
   Support of the random variable X with PDF f_X is
   $$\text{supp}(X) = \{x : f_X(x) > 0\}$$
   - $\text{supp}(X)$ contains intervals in which X can fall with positive probability.

10. **Continuous Uniform distribution:**
    - $X \sim \text{Uniform}[a, b]$
    - PDF:
    $$f_X(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$
    - CDF:
    $$F_X(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$
11. **Exponential distribution:**
    - $X \sim \text{Exp}(\lambda)$
    - PDF:
    $$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$
    - CDF:
    $$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$
12. **Normal distribution:**
    - $X \sim \text{Normal}[\mu, \sigma^2]$
    - PDF:
    $$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \qquad -\infty < x < \infty$$
    - CDF:
    $$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$
    - CDF has no closed form expression.
    - Standard normal: $Z = \text{Normal}(0,\ 1)$
      - PDF: $f_Z(z) = \dfrac{1}{\sqrt{2\pi}} \exp\left(\dfrac{-z^2}{2}\right) \qquad -\infty < z < \infty$
13. **Standardization:**
    If $X \sim \text{Normal}(\mu, \sigma^2)$, then
    $$\frac{X-\mu}{\sigma} = Z \sim \text{Normal}(0, 1)$$
14. To compute the probabilities of the normal distribution, convert probability computation to that of a standard normal.

15. **Functions of continuous random variable:**
Suppose X is a continuous random variable with CDF F_X and PDF f_X and suppose $g : \mathbb{R} \to \mathbb{R}$ is a (reasonable) function. Then, $Y = g(X)$ is a random variable with CDF F_Y determined as follows:

   - $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \in \{x : g(x) \leq y\})$
   - To evaluate the above probability
     - Convert the subset $A_y = \{x : g(x) \leq y\}$ into intervals in real line.
     - Find the probability that X falls in those intervals.
     - $F_Y(y) = P(X \in A_Y) = \int_{A_Y} f_X(x)dx$
   - If F_Y has no jumps, you may be able to differentiate and find a PDF.

16. **Theorem: Monotonic differentiable function**
Suppose X is a continuous random variable with PDF f_X . Let $g(x)$ be monotonic for $x \in \text{supp}(X)$ with derivative $g'(x) = \dfrac{dg(x)}{dx}$. Then, the PDF of $Y = g(X)$ is

$$f_Y(y) = \frac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y))$$

   - **Translation:** $Y = X + a$
$$f_Y(y) = f_X(y - a)$$
   - **Scaling:** $Y = aX$
$$f_Y(y) = \frac{1}{|a|} f_X(y/a)$$
   - **Affine:** $Y = aX + b$
$$f_Y(y) = \frac{1}{|a|} f_X((y-b)/a)$$
   - Affine transformation of a normal random variable is normal.

17. **Expected value of function of continuous random variable:**
Let X be a continuous random variable with density $f_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function. The expected value of $g(X)$, denoted $E[g(X)]$, is given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

whenever the above integral exists.

   - The integral may diverge to $\pm\infty$ or may not exist in some cases.

18. **Expected value (mean) of a continuous random variable:**
Mean, denoted $E[X]$ or μ_X or simply μ is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

19. **Variance of a continuous random variable:**
Variance, denoted Var$[X]$ or σ_X^2 or simply σ^2 is given by

$$\text{Var}(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

- Variance is a measure of spread of X about its mean.
- $\text{Var}(X) = E[X^2] - E[X]^2$

X	$E[X]$	$\text{Var}(X)$
Uniform$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal(μ, σ^2)	μ	σ^2

20. **Markov's inequality:**
If X is a continuous random variable with mean μ and non-negative supp(X) (i.e. $P(X < 0) = 0$), then

$$P(X > c) \leq \frac{\mu}{c}$$

21. **Chebyshev's inequality:**
If X is a continuous random variable with mean μ and variance σ^2, then

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Statistics for Data Science - 2

Week 6 Notes

1. **Marginal density:** Let (X, Y) be jointly distributed where X is discrete with range T_X and PMF $p_X(x)$.
   For each $x \in T_X$, we have a continuous random variable Y_x with density $f_{Y_x}(y)$.
   $f_{Y_x}(y)$: conditional density of Y given $X = x$, denoted $f_{Y|X=x}(y)$.
   - Marginal density of Y
     - $f_Y(y) = \sum\limits_{x \in T_X} p_X(x) f_{Y|X=x}(y)$
2. **Conditional probability of discrete given continuous:** Suppose X and Y are jointly distributed with $X \in T_X$ being discrete with PMF $p_X(x)$ and conditional densities $f_{Y|X=x}(y)$ for $x \in T_X$. The conditional probability of X given $Y = y_0 \in \text{supp}(Y)$ is defined as
   - $P(X = x \mid Y = y_0) = \dfrac{p_X(x) f_{Y|X=x}(y_0)}{f_Y(y_0)}$
3. **Joint density:** A function $f(x, y)$ is said to be a joint density function if
   - $f(x, y) \geq 0$, i.e. f is non-negative.
   - $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y) dx dy = 1$
4. **2D uniform distribution:** Fix some (reasonable) region D in $\mathbb{R}^2$ with total area $|D|$. We say that $(X, Y) \sim \text{Uniform}(D)$ if they have the joint density
   $$f_{XY}(x, y) = \begin{cases} \frac{1}{|D|} & (x, y) \in D \\ 0 & \text{otherwise} \end{cases}$$
5. **Marginal density:** Suppose (X, Y) have joint density $f_{XY}(x, y)$. Then,
   - X has the marginal density $f_X(x) = \int\limits_{y=-\infty}^{y=\infty} f_{XY}(x, y) dy$.
   - Y has the marginal density $f_Y(y) = \int\limits_{x=-\infty}^{x=\infty} f_{XY}(x, y) dx$.
     - In general the marginals do not determine joint density.
6. **Independence:** (X, Y) with joint density $f_{XY}(x, y)$ are independent if

- $f_{XY}(x, y) = f_X(x) f_Y(y)$
  - If independent, the marginals determine the joint density.

7. **Conditional density:** Let (X, Y) be random variables with joint density $f_{XY}(x, y)$. Let $f_X(x)$ and $f_Y(y)$ be the marginal densities.

   - For a such that $f_X(a) > 0$, the conditional density of Y given $X = a$, denoted as $f_{Y|X=a}(y)$, is defined as

   $$f_{Y|X=a}(y) = \frac{f_{XY}(a, y)}{f_X(a)}$$

   - For b such that $f_Y(b) > 0$, the conditional density of X given $Y = b$, denoted as $f_{X|Y=b}(x)$, is defined as

   $$f_{X|Y=b}(x) = \frac{f_{XY}(x, b)}{f_Y(b)}$$

8. **Properties of conditional density:** Joint = Marginal $\times$ Conditional, for $x = a$ and $y = b$ such that $f_X(a) > 0$ and $f_Y(b) > 0$.

   - $f_{XY}(a, b) = f_X(a) f_{Y|X=a}(b) = f_Y(b) f_{X|Y=b}(a)$

Statistics for Data Science - 2

Important results

Discrete random variables:

Distribution	PMF ($f_X(k)$)	CDF ($F_X(x)$)	$E[X]$	Var(X)
Uniform(A) $A = \{a, a+1, \ldots, b\}$	$\frac{1}{n}, \quad x = k$ $n = b - a + 1$ $k = a, a+1, \ldots, b$	$\begin{cases} 0 & x < 0 \\ \frac{k-a+1}{n} & k \leq x < k+1 \\ & k = a, a+1, \ldots, b-1, b \\ 1 & x \geq n \end{cases}$	$\frac{a+b}{2}$	$\frac{n^2-1}{12}$
Bernoulli(p)	$\begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$	$\begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$	p	$p(1-p)$
Binomial(n, p)	${}^nC_k p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$	$\begin{cases} 0 & x < 0 \\ \sum_{i=0}^{k} {}^nC_i p^i (1-p)^{n-i} & k \leq x < k+1 \\ & k = 0, 1, \ldots, n \\ 1 & x \geq n \end{cases}$	np	$np(1-p)$
Geometric(p)	$(1-p)^{k-1}p$, $k = 1, \ldots, \infty$	$\begin{cases} 0 & x < 0 \\ 1-(1-p)^k & k \leq x < k+1 \\ & k = 1, \ldots, \infty \end{cases}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Poisson(λ)	$\dfrac{e^{-\lambda}\lambda^k}{k!}$, $k = 0, 1, \ldots, \infty$	$\begin{cases} 0 & x < 0 \\ e^{-\lambda} \sum_{i=0}^{k} \dfrac{\lambda^i}{i!} & k \leq x < k+1 \\ & k = 0, 1, \ldots, \infty \end{cases}$	λ	λ

Continuous random variables:

Distribution	PDF ($f_X(k)$)	CDF ($F_X(x)$)	$E[X]$	$\text{Var}(X)$
Uniform$[a,b]$	$\frac{1}{b-a}$, $a \leq x \leq b$	$\begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$\lambda e^{-\lambda x}$, $x > 0$	$\begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$, $-\infty < x < \infty$	No closed form	μ	σ^2
Gamma(α, β)	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $x > 0$		$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $0 < x < 1$		$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

1. **Markov's inequality:** Let X be a discrete random variable taking non-negative values with a finite mean μ. Then,
$$P(X \geq c) \leq \frac{\mu}{c}$$

2. **Chebyshev's inequality:** Let X be a discrete random variable with a finite mean μ and a finite variance σ^2. Then,
$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

3. **Weak Law of Large numbers:** Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
Define sample mean $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$. Then,
$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$

4. **Using CLT to approximate probability:** Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
Define $Y = X_1 + X_2 + \ldots + X_n$. Then,
$$\frac{Y - n\mu}{\sqrt{n}\sigma} \approx \text{Normal}(0, 1).$$

- **Test for mean**
  Case (1): When population variance σ^2 is known (z-test)

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu = \mu_0$	$\mu > \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\overline{X} > c$
left-tailed	$\mu = \mu_0$	$\mu < \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\overline{X} < c$
two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert\overline{X} - \mu_0\rvert > c$

Case (2): When population variance σ^2 is unknown (t-test)

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu = \mu_0$	$\mu > \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\overline{X} > c$
left-tailed	$\mu = \mu_0$	$\mu < \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\overline{X} < c$
two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert\overline{X} - \mu_0\rvert > c$

- **χ^2-test for variance:**

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\sigma = \sigma_0$	$\sigma > \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 > c^2$
left-tailed	$\sigma = \sigma_0$	$\sigma < \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 < c^2$
two-tailed	$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 > c^2$ where $\dfrac{\alpha}{2} = P(S^2 > c^2)$ or $S^2 < c^2$ where $\dfrac{\alpha}{2} = P(S^2 < c^2)$

- **Two samples z-test for means:**

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$T = \overline{X} - \overline{Y}$ $\overline{X} - \overline{Y} \sim \text{Normal}\left(0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ if H_0 is true	$\overline{X} - \overline{Y} > c$
left-tailed	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$T = \overline{Y} - \overline{X}$ $\overline{Y} - \overline{X} \sim \text{Normal}\left(0, \dfrac{\sigma_2^2}{n_2} + \dfrac{\sigma_1^2}{n_1}\right)$ if H_0 is true	$\overline{Y} - \overline{X} > c$
two-tailed	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \overline{X} - \overline{Y}$ $\overline{X} - \overline{Y} \sim \text{Normal}\left(0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ if H_0 is true	$\lvert\overline{X} - \overline{Y}\rvert > c$

- **Two samples F-test for variances**

Test	H_0	H_A	Test statistic	Rejection region
one-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 > \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} > 1 + c$
one-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 < \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} < 1 - c$
two-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} > 1 + c_R$ where $\dfrac{\alpha}{2} = P(T > 1 + c_R)$ or $\dfrac{S_1^2}{S_2^2} < 1 - c_L$ where $\dfrac{\alpha}{2} = P(T < 1 - c_L)$

- **χ^2-test for goodness of fit:**
  H_0 : Samples are i.i.d X, $\quad H_A$: Samples are not i.i.d X

  $$\text{Test statistic: } T = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{k-1}$$

  Test: Reject H_0 if $T > c$.

- **Test for independence:**
  H_0 : Joint PMF is product of marginals, H_A : Joint PMF is not product of marginals

  $$\text{Test statistic: } T = \sum_{i,j} \frac{(y_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{dof}$$

  where dof = (number of rows-1) $\times$ (number of columns-1)
  y_{ij} = product of marginals for (i, j)
  np_{ij} = expected, if independent

  Test: Reject H_0 if $T > c$.

Statistics for Data Science - 2

Week 7 Notes

Statistics from samples and Limit theorems

1. **Empirical distribution:**
Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. Let $\#(X_i = t)$ denote the number of times t occurs in the samples. The empirical distribution is the discrete distribution with PMF
$$p(t) = \frac{\#(X_i = t)}{n}$$
   - The empirical distribution is random because it depends on the actual sample instances.
   - **Descriptive statistics:** Properties of empirical distribution. Examples :
     - Mean of the distribution
     - Variance of the distribution
     - Probability of an event
   - As number of samples increases, the properties of empirical distribution should become close to that of the original distribution.

2. **Sample mean:**
Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. The sample mean, denoted $\overline{X}$, is defined to be the random variable
$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$
   - Given a sampling $x_1, \ldots, x_n$ the value taken by the sample mean $\overline{X}$ is $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$. Often, $\overline{X}$ and $\overline{x}$ are both called sample mean.

3. **Expected value and variance of sample mean:**
Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples whose distribution has a finite mean μ and variance σ^2. The sample mean $\overline{X}$ has expected value and variance given by
$$E[\overline{X}] = \mu, \qquad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$
   - Expected value of sample mean equals the expected value or mean of the distribution.
   - Variance of sample mean decreases with n.

4. **Sample variance:**
   Let $X_1, X_2, \ldots, X_n \sim X$ be i.i.d. samples. The sample variance, denoted S^2, is defined to be the random variable

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1},$$

   where $\overline{X}$ is the sample mean.

5. **Expected value of sample variance:**
   Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples whose distribution has a finite variance σ^2. The sample variance $S^2 = \dfrac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$ has expected value given by

$$E[S^2] = \sigma^2$$

   - Values of sample variance, on average, give the variance of distribution.
   - Variance of sample variance will decrease with number of samples (in most cases).
   - As n increases, sample variance takes values close to distribution variance.

6. **Sample proportion:**
   The sample proportion of A, denoted $S(A)$, is defined as

$$S(A) = \frac{\text{number of } X_i \text{ for which } A \text{ is true}}{n}$$

   - As n increases, values of $S(A)$ will be close to $P(A)$.
   - Mean of $S(A)$ equals $P(A)$.
   - Variance of $S(A)$ tends to 0.

7. **Weak law of large numbers:**
   Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
   Define sample mean $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$. Then,

$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$

8. **Chernoff inequality:**
   Let X be a random variable such that $E[X] = 0$, then

$$P(X > t) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}}, \quad \lambda > 0$$

9. **Moment generating function (MGF):**
Let X be a zero-mean random variable ($E[X] = 0$). The MGF of X, denoted $M_X(\lambda)$, is a function from $\mathbb{R}$ to $\mathbb{R}$ defined as

$$M_X(\lambda) = E[e^{\lambda X}]$$

   -
$$\begin{aligned} M_X(\lambda) &= E[e^{\lambda X}] \\ &= E[1 + \lambda X + \frac{\lambda^2 X^2}{2!} + \frac{\lambda^3 X^3}{3!} + \ldots] \\ &= 1 + \lambda E[X] + \frac{\lambda^2}{2!} E[X^2] + \frac{\lambda^3}{3!} E[X^3] + \ldots \end{aligned}$$

     That is coefficient of $\frac{\lambda^k}{k!}$ in the MGF of X gives the kth moment of X.
   - If $X \sim \text{Normal}(0, \sigma^2)$ then, $M_X(\lambda) = e^{\lambda^2 \sigma^2/2}$
   - Let $X_1, X_2, \ldots, X_n \sim$ i.i.d. X and let $S = X_1 + X_2 + \ldots + X_n$, then

$$M_S(\lambda) = (E[e^{\lambda X}])^n = [M_X(\lambda)]^n$$

     It implies that MGF of sum of independent random variables is product of the individual MGFs.

10. **Central limit theorem:** Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$. Define $Y = X_1 + X_2 + \ldots + X_n$. Then,

$$\frac{Y - n\mu}{\sqrt{n}\sigma} \approx \text{Normal}(0, 1).$$

11. **Gamma distribution:**
$X \sim \text{Gamma}(\alpha, \beta)$ if PDF $f_x(x) \propto x^{\alpha-1} e^{-\beta x}, \quad x > 0$

   - $\alpha > 0$ is a shape parameter.
   - $\beta > 0$ is a rate parameter.
   - $\theta = \frac{1}{\beta}$ is a scale parameter.
   - Mean, $E[X] = \frac{\alpha}{\beta}$
   - Variance, $\text{Var}(X) = \frac{\alpha}{\beta^2}$

12. **Beta distribution:**
$X \sim \text{Beta}(\alpha, \beta)$ if PDF $f_x(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$
    - $\alpha > 0, \beta > 0$ are the shape parameters.
    - Mean, $E[X] = \dfrac{\alpha}{\alpha+\beta}$
    - Variance, $\text{Var}(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

13. **Cauchy distribution:**
$X \sim \text{Cauchy}(\theta, \alpha^2)$ if PDF $f_x(x) \propto \dfrac{1}{\pi}\dfrac{\alpha}{\alpha^2+(x-\theta)^2}$
    - θ is a location parameter.
    - $\alpha > 0$ is a scale parameter.
    - Mean and variance are undefined.

14. **Some important results:**
    - Let $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$ are independent and let $Y = a_1X_1 + a_2X_2 + \ldots a_nX_n$, then
    $$Y \sim \text{Normal}(\mu, \sigma^2)$$
    where $\mu = a_1\mu_1 + a_2\mu_2 + \ldots a_n\mu_n$ and $\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots a_n^2\sigma_n^2$
    That is linear combinations of i.i.d. normal distributions is again a normal distribution.
    - Sum of n i.i.d. $\text{Exp}(\beta)$ is $\text{Gamma}(n, \beta)$.
    - Square of $\text{Normal}(0, \sigma^2)$ is $\text{Gamma}\left(\dfrac{1}{2}, \dfrac{1}{2\sigma^2}\right)$.
    - Suppose $X, Y \sim$ i.i.d. $\text{Normal}(0, \sigma^2)$. Then, $\dfrac{X}{Y} \sim \text{Cauchy}(0, 1)$.
    - Suppose $X \sim \text{Gamma}(\alpha, k), Y \sim \text{Gamma}(\beta, k)$ are independent random variables, then $\dfrac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$.
    - Sum of n independent $\text{Gamma}(\alpha, \beta)$ is $\text{Gamma}(n\alpha, \beta)$.
    - If $X_1, X_2, \ldots, X_n \sim$ i.i.d. $\text{Normal}(0, \sigma^2)$, then $X_1^2 + X_2^2 + \ldots + X_n^2 \sim \text{Gamma}\left(\dfrac{n}{2}, \dfrac{1}{2\sigma^2}\right)$.

- Gamma$\left(\frac{n}{2}, \frac{1}{2}\right)$ is called Chi-square distribution with n degrees of freedom, denoted χ^2_n.

- Suppose $X_1, X_2, \ldots, X_n \sim$ i.i.d. Normal(μ, σ^2). Suppose that $\overline{X}$ and S^2 denote the sample mean and sample variance, respectively, then
  (i) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
  (ii) $\overline{X}$ and S^2 are independent.

Statistics for Data Science - 2

Week 8 notes

- Let $X_1, \ldots, X_n \sim$ i.i.d.X, where X has the distribution described by parameters $\theta_1, \theta_2, \ldots$.
  - The parameters θ_i are unknown but a fixed constant.
  - Define the estimator for θ as the function of the samples: $\hat{\theta}(X_1, \ldots, X_n)$.

  Note:

  1. θ is an unknown parameter.
  2. $\hat{\theta}$ is a function of n random variables.

  Remark: Infinite number of estimators are possible for a parameter of a distribution.

- Estimation error: $\hat{\theta}(X_1, \ldots, X_n) - \theta$ is a random variable.
  - We expect the estimator random variable $\hat{\theta}(X_1, \ldots, X_n)$ to take values around the actual value of the parameter θ. So, the random variable 'Error' should take values close to 0.
  - Mathematically, it is expressed as $P(\mid \text{Error} \mid > \delta)$ should be small.
  - Chebyshev bound on error: $P(\mid \text{Error} - E[\text{Error}] \mid > \delta) \leq \dfrac{\text{Var(Error)}}{\delta^2}$.
  - Good design: $P(\mid \text{Error} \mid > \delta)$ will fall with n.

- Good design principles:
  1. Error should be close to or equal to 0.
  2. Var(Error) $\rightarrow 0$ with n.

- Bias: The bias of the estimator $\hat{\theta}$

  for a parameter θ, denoted Bias$(\hat{\theta}, \theta)$ is defined as

$$\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$$

  1. Bias is the expected value of Error.
  2. An estimator with bias equal to 0 is said to be an unbiased estimator.

- Risk: The (squared-error) risk of the estimator $\hat{\theta}$ for a parameter θ, denoted Risk$(\hat{\theta}, \theta)$, is defined as

$$\text{Risk}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$$

1. Risk is the expected value of "squared error" and is also called mean squared error (MSE) often.
2. Squared-error risk is the second moment of Error.

- Variance of estimator:

$$\text{Variance}(\hat{\theta}) = E[(\hat{\theta} - E[\theta])^2]$$

 $\text{Var}(\text{Error}) = \text{Var}(\hat{\theta})$

- Bias-Variance tradeoff: The risk of the estimator satisfies the following relationship:

$$\text{Risk}(\hat{\theta}, \theta) = \text{Bias}(\hat{\theta}, \theta)^2 + \text{Variance}(\hat{\theta})$$

- Estimator design approach:

1. Method of moments
   (a) Sample moments: $M_k(X_1, \ldots, X_n) = \dfrac{1}{n} \sum_{i=1}^{n} X_i^k$
   (b) M_k is a random variable, and m_k is the value taken by it in one sampling instance. We expect that M_k will take values around $E[X^k]$
   (c) Procedure:
      - Equate sample moments to expression for moments in terms of unknown parameters.
      - Solve for the unknown parameters.
   (d) One parameter θ usually needs one moment
      - Sample moment: m_1
      - Distribution moment: $E[X] = f(\theta)$
      - Solve for θ from $f(\theta) = m_1$ in terms of m_1.
      - $\hat{\theta}$: replace m_1 by M_1 in above solution.
   (e) Two parameters θ_1, θ_2 usually needs two moments.
      - Sample moments: m_1, m_2
      - Distribution moment: $E[X] = f(\theta_1, \theta_2), E[X^2] = g(\theta_1, \theta_2)$
      - Solve for θ_1, θ_2 from $f(\theta_1, \theta_2) = m_1, g(\theta_1, \theta_2) = m_2$ in terms of m_1, m_2.
      - $\hat{\theta}$: replace m_1 by M_1 and m_2 by M_2in above solution.
2. Maximum Likelihood estimators
   (a) Likelihood of i.i.d. samples: Likelihood of a sampling $x_1, x_2, \ldots, x_n$, denoted $L(x_1, x_2, \ldots, x_n)$

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

      - Likelihood $L(x_1, x_2, \ldots, x_n)$ is a function of parameters.

– Maximum likelihood (ML) estimation

$$\theta_1^*, \theta_2^*, \ldots = arg \max_{\theta_1,\theta_2,\ldots} \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

We find parameters that maximize likelihood for a given set of samples.

- Properties of estimators:

1. Consistency of estimators: If an estimator satisfies the following requirement, it is said to be consistent. Technically, it is called convergence in probability.
   $P(|\text{ Error }| > \delta) \to 0$ as $n \to \infty$ for any $\delta > 0$.
2. To compare the estimators, use mean squared error (MSE).

- Confidence interval:

$$X_1, \ldots, X_n \sim \text{iid } X, \mu = E[X]$$

Estimator: $\hat{\mu} = \dfrac{X_1 + \ldots + X_n}{n}$

– Suppose $P(|\hat{\mu} - \mu| < \alpha) = \beta$, where α is a small fraction and β is a large fraction.

– $\hat{\mu}$ in one sampling instance: estimate with margin of error $(100\alpha)\%$ at confidence level $(100\beta)\%$.

1. Normal samples with known variance: $X_1, \ldots, X_n \sim$ iid Normal$(\mu, \sigma^2), \sigma^2$ known.
   Estimator: $\hat{\mu} = \dfrac{X_1 + \ldots + X_n}{n}$
   $\hat{\mu} \sim$ i.i.d. Normal$(\mu, \frac{\sigma^2}{n}), Z = \dfrac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim$ Normal$(0, 1)$

$$\begin{aligned} P(|\hat{\mu} - \mu| < \alpha) &= \beta \\ \implies P\left(\left|\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}\right| < \frac{\alpha}{\sigma/\sqrt{n}}\right) &= \beta \\ \implies P\left(|\text{ Normal}(0,1)\,| < \frac{\alpha}{\sigma/\sqrt{n}}\right) &= \beta \end{aligned}$$

2. Normal samples with unknown variance: $X_1, \ldots, X_n \sim$ iid Normal$(\mu, \sigma^2), \sigma^2$ unknown.
   Sampling instance: $x_1, \ldots, x_n$.
   Estimated mean and variance: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i, \hat{\sigma}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$\hat{\mu} \sim$ i.i.d. Normal$(\mu, \frac{\sigma^2}{n}), Z = \dfrac{\hat{\mu} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$$\begin{aligned} P(\mid \hat{\mu} - \mu \mid < \alpha) =& \beta \\ \implies P\left(\left| \frac{\hat{\mu} - \mu}{S/\sqrt{n}} \right| < \frac{\alpha}{\hat{\alpha}/\sqrt{n}} \right) =& \beta \\ \implies P\left(\mid \text{Normal}(0,1) \mid < \frac{\alpha}{\hat{\alpha}/\sqrt{n}} \right) =& \beta \end{aligned}$$

3. <u>If samples are not normal</u>: Use CLT to argue that sample mean will have a normal distribution

Statistics for Data Science - 2

Week 9 Notes

1. **Parameter estimation:** Let $X_1, \ldots, X_n \sim$ iid X, parameter Θ
   Prior distribution of Θ: $\Theta \sim f_\Theta(\theta)$
   Samples: $x_1, \ldots, x_n$, notation $S = (X_1 = x_1, \ldots X_n = x_n)$
   Bayes' rule: posterior $\propto$ likelihood $\times$ prior

$$P(\Theta = \theta \mid S) = P(S \mid \Theta = \theta) f_\Theta(\theta) / P(S)$$

   In case of discrete: $P(S) = \sum\limits_{\theta} P(S \mid \Theta = \theta) f_\Theta(\theta)$
   In case of continuous: $P(S) = \int\limits_{\theta} P(S \mid \Theta = \theta) f_\Theta(\theta)\ d\theta$
   Posterior mode: $\hat{\theta} = \arg\max_\theta P(S \mid \Theta = \theta) f_\Theta(\theta)$
   Posterior mean: $\mathrm{E}[\Theta \mid S]$, mean of posterior distribution.

2. **Bernoulli(p) samples with uniform prior:** $X_1, \ldots, X_n \sim$ iid Bernoulli($\mathbf{p}$)
   Prior $\mathbf{p} \sim$ Uniform[0, 1]
   Samples: $x_1, \ldots, x_n$
   Posterior: $\mathbf{p} |\ (X_1 = x_1, \ldots X_n = x_n)$
   Posterior density $\propto P(X_1 = x_1, \ldots X_n = x_n \mid \mathbf{p} = p) \times f_{\mathbf{p}}(p)$
   Posterior density $\propto p^w (1-p)^{n-w}$
   $\Rightarrow$ Posterior density: Beta$(w+1, n-w+1)$
   Posterior mean: $\hat{p} = \dfrac{X_1 + X_2 + \ldots + X_n + 1}{n+2}$

3. **Bernoulli(p) samples with beta prior:** $X_1, \ldots, X_n \sim$ iid Bernoulli($\mathbf{p}$)
   Prior $\mathbf{p} \sim$ Beta(α, β)
   $\Rightarrow f_{\mathbf{p}}(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$
   Samples: $x_1, \ldots, x_n$
   Posterior: $\mathbf{p} |\ (X_1 = x_1, \ldots X_n = x_n)$
   Posterior density $\propto P(X_1 = x_1, \ldots X_n = x_n \mid \mathbf{p} = p) \times f_{\mathbf{p}}(p)$
   Posterior density $\propto p^{w+\alpha-1}(1-p)^{n-w+\beta-1}$

   $\Rightarrow$ Posterior density: Beta$(w+\alpha, n-w+\beta)$
   Posterior mean: $\hat{p} = \dfrac{X_1 + X_2 + \ldots + X_n + \alpha}{n+\alpha+\beta}$

4. **Normal samples with unknown mean and known variance:** $X_1, \ldots, X_n \sim$ iid Normal(M, σ^2)

Prior M $\sim$ Normal(μ_0, σ_0^2)
$\Rightarrow f_M(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$
Samples: $x_1, \ldots, x_n$, Sample mean: $\overline{x} = (x_1 + \ldots + x_n)/n$
Posterior: M$|$ $(X_1 = x_1, \ldots X_n = x_n)$
Posterior density $\propto f(X_1 = x_1, \ldots X_n = x_n \mid M = \mu) \times f_M(\mu)$
Posterior density $\propto \exp(-\frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{2\sigma_0^2})\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$
$\Rightarrow$ Posterior density: Normal
Posterior mean: $\hat{\mu} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}\dfrac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0\dfrac{\sigma^2}{n\sigma_0^2 + \sigma^2}$

5. **Geometric(p) samples with Uniform[0, 1] prior:** $X_1, \ldots, X_n \sim$ iid Geometric($\mathbf{p}$)
   Prior $\mathbf{p} \sim$ Uniform[0, 1]
   Samples: $x_1, \ldots, x_n$
   Posterior: $\mathbf{p}|$ $(X_1 = x_1, \ldots X_n = x_n)$
   Posterior density $\propto P(X_1 = x_1, \ldots X_n = x_n \mid \mathbf{p} = p) \times f_{\mathbf{p}}(p)$
   Posterior density $\propto p^n(1-p)^{x_1+\ldots+x_n-n}$

   $\Rightarrow$ Posterior density: Beta$(n+1, x_1 + \ldots + x_n - n + 1)$
   Posterior mean: $\hat{p} = \dfrac{n+1}{X_1 + \ldots + X_n + 2}$

6. **Poisson(λ) samples with gamma prior:** $X_1, \ldots, X_n \sim$ iid Poisson(Λ)
   Prior $\Lambda \sim$ Gamma(α, β)
   $\Rightarrow f_\Lambda(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$
   Samples: $x_1, \ldots, x_n$
   Posterior: $\Lambda \mid (X_1 = x_1, \ldots X_n = x_n)$
   Posterior density $\propto P(X_1 = x_1, \ldots X_n = x_n \mid \Lambda = \lambda) \times f_\Lambda(\lambda)$
   Posterior density $\propto e^{-n\lambda}\lambda^{x_1+\ldots+x_n}\lambda^{\alpha-1}e^{-\beta\lambda}$

   $\Rightarrow$ Posterior density: Gamma$(x_1 + \ldots + x_n + \alpha, \beta + n)$

   Posterior mean: $\hat{\lambda} = \dfrac{X_1 + X_2 + \ldots + X_n + \alpha}{n + \beta}$

Statistics for Data Science - 2

Week 10 Notes

Hypothesis testing

1. **Null hypothesis:**
The null hypothesis is a kind of hypothesis which explains the population parameter whose purpose is to test the validity of the given experimental data. It is denoted by H_0. The null hypothesis is a default hypothesis that is assumed to remain possibly true.

2. **Alternative hypothesis:**
The alternative hypothesis is a statement used in statistical inference experiment. It is contradictory to the null hypothesis and denoted by H_A or H_1.

3. **Test statistic:**
A test statistic is numerical quantity computed from values in a sample used in statistical hypothesis testing.

4. **Type I error:**
A type I error is a kind of fault that occurs during the hypothesis testing process when a null hypothesis is rejected, even though it is true.

5. **Type II error:**
A type II error is a kind of fault that occurs during the hypothesis testing process when a null hypothesis is accepted, even though it is not true (H_A is true).

6. **Significance level (Size):**
Significance level (also called size) of a test, denoted α, is the probability of type I error.
$$\alpha = P(\text{Type I error})$$

7. $\beta = P(\text{Type II error})$

8. **Power of a test:**
Power $= 1 - \beta$

9. **Types of hypothesis:**

   (a) **Simple hypothesis:** A hypothesis that completely specifies the distribution of the samples is called a simple hypothesis.

   (b) **Composite hypothesis:** A hypothesis that does not completely specify the distribution of the samples is called a composite hypothesis.

10. **Standard testing method: z-test:**
Consider a sample $X_1, X_2, \ldots, X_n \sim$ i.i.d. X.

- Test statistic, denoted T, is some function of the samples. For example: sample mean $\overline{X}$
- Acceptance and rejection regions are specified through T.

(a) **Right-tailed z-test:**

- $H_0 : \mu = \mu_0, \quad H_A : \mu > \mu_0$
- Test: reject H_0 if $T > c$.
- Significance level α depends on c and the distribution of $T|H_0$.
- $\alpha = P(T > c|H_0)$
- Fix α and find c.

(b) **Left-tailed z-test:**

- $H_0 : \mu = \mu_0, \quad H_A : \mu < \mu_0$
- Test: reject H_0 if $T < c$.
- Significance level α depends on c and the distribution of $T|H_0$.
- $\alpha = P(T < c|H_0)$
- Fix α and find c.

(c) **two-tailed z-test:**

- $H_0 : \mu = \mu_0, \quad H_A : \mu \neq \mu_0$
- Test: reject H_0 if $|T| > c$.
- Significance level α depends on c and the distribution of $T|H_0$.
- $\alpha = P(|T| > c|H_0)$
- Fix α and find c.

Note: In the test for mean (σ^2 known), $T = \overline{X}$ and when null is true, $\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim$ Normal$(0, 1)$.

11. ***P*-value:**
Suppose the test statistic $T = t$ in one sampling. The lowest significance level α at which the null will be rejected for $T = t$ is said to be the P-value of the sampling.

Statistics for Data Science - 2

Week 11 Notes

t-test, χ^2-test, two samples z/F-test

1. **Normal samples and statistics:** Consider the samples $X_1, \ldots, X_n \sim$ iid Normal(μ, σ^2).
The sample mean, $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$
The sample variance, $S^2 = \dfrac{1}{n-1}[(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2]$
$E[\overline{X}] = \mu$, $E[S^2] = \sigma^2$
   - $\overline{X} \sim$ Normal$(\mu, \sigma^2/n)$
   - $\dfrac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$, chi-squared distribution with $n-1$ degrees of freedom.
   - $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$, t-distribution with $n-1$ degrees of freedom.

2. **t-test for mean (Variance unknown)**
Consider the samples $X_1, \ldots, X_n \sim$ iid Normal(μ, σ^2), σ^2 unknown. Following are the three different possibilities:
   - The null and alternative hypothesis are:
   $$H_0 : \mu = \mu_0$$
   $$H_A : \mu > \mu_0$$
   Test Statistic: $T = \overline{X}$
   Test: Reject H_0, if $T > c$
   Given H_0, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
   $$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(T > c \mid \mu = \mu_0)\\ =& P\left(t_{n-1} > \frac{c - \mu_0}{s/\sqrt{n}}\right) = 1 - F_{t_{n-1}}\left(\frac{c - \mu_0}{s/\sqrt{n}}\right)\\ \implies c =& \frac{s}{\sqrt{n}}F^{-1}_{t_{n-1}}(1 - \alpha) + \mu_0\end{aligned}$$
   Note: $F_{t_{n-1}}$ is the CDF of t-distribution with $n-1$ degrees of freedom.
   - The null and alternative hypothesis are:
   $$H_0 : \mu = \mu_0$$

$$H_A : \mu < \mu_0$$

Test Statistic: $T = \overline{X}$
Test: Reject H_0, if $T < c$
Given H_0, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(T < c \mid \mu = \mu_0)\\ =& P\left(t_{n-1} < \frac{c - \mu_0}{s/\sqrt{n}}\right) = F_{t_{n-1}}\left(\frac{c-\mu_0}{s/\sqrt{n}}\right)\\ \implies c =& \frac{s}{\sqrt{n}} F^{-1}_{t_{n-1}}(\alpha) + \mu_0\end{aligned}$$

Note: $F_{t_{n-1}}$ is the CDF of t-distribution with $n - 1$ degrees of freedom.

- The null and alternative hypothesis are:

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0$$

Test Statistic: $T = \overline{X} - \mu$
Test: Reject H_0, if $\mid \overline{X} - \mu \mid > c$
Given H_0, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(\mid \overline{X} - \mu \mid > c \mid \mu = \mu_0)\\ =& P\left(\mid t_{n-1} \mid > \frac{c}{s/\sqrt{n}}\right) = 2F_{t_{n-1}}\left(\frac{-c}{s/\sqrt{n}}\right)\\ \implies c =& \frac{-s}{\sqrt{n}} F^{-1}_{t_{n-1}}(\alpha/2)\end{aligned}$$

Note: $F_{t_{n-1}}$ is the CDF of t-distribution with $n - 1$ degrees of freedom.

3. **χ^2-test for variance**
Consider the samples $X_1, \ldots, X_n \sim$ iid Normal(μ, σ^2), σ^2 unknown. Following are the three different possibilities:

- The null and alternative hypothesis are:

$$H_0 : \sigma = \sigma_0$$

$$H_A : \sigma > \sigma_0$$

Test Statistic: S^2
Test: Reject H_0, if $S^2 > c^2$
Given H_0, $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(S^2 > c^2 \mid \sigma = \sigma_0)\\ =& P\left(\chi^2_{n-1} > \frac{(n-1)}{\sigma_0^2}c^2\right) = 1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)}{\sigma_0^2}c^2\right)\end{aligned}$$

Note: $F_{\chi^2_{n-1}}$ is the CDF of chi-distribution with $n-1$ degrees of freedom.

- The null and alternative hypothesis are:

$$H_0 : \sigma = \sigma_0$$
$$H_A : \sigma < \sigma_0$$

Test Statistic: S^2
Test: Reject H_0, if $S^2 < c^2$
Given H_0, $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(S^2 < c^2 \mid \sigma = \sigma_0)\\ =& P\left(\chi^2_{n-1} < \frac{(n-1)}{\sigma_0^2}c^2\right) = F_{\chi^2_{n-1}}\left(\frac{(n-1)}{\sigma_0^2}c^2\right)\end{aligned}$$

Note: $F_{\chi^2_{n-1}}$ is the CDF of chi-distribution with $n-1$ degrees of freedom.

- The null and alternative hypothesis are:

$$H_0 : \sigma = \sigma_0$$
$$H_A : \sigma \neq \sigma_0$$

Test Statistic: S^2
Test: Reject H_0, if $S^2 < c^2$ or $S^2 > c^2$
Given H_0, $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$

$$\frac{\alpha}{2} = P(S^2 < c^2 \mid H_0) = P(S^2 > c^2 \mid H_0)$$

Note: $F_{\chi^2_{n-1}}$ is the CDF of chi-distribution with $n-1$ degrees of freedom.

4. **Two samples z-test (known variances)**

Let $X_1, \ldots, X_{n_1} \sim$ iid Normal(μ_1, σ_1^2)
and $Y_1, \ldots, Y_{n_2} \sim$ iid Normal(μ_2, σ_2^2)
Following are the three different possibilities:

- The null and alternative hypothesis are:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

Test Statistic: $T = \overline{X} - \overline{Y}$
Test: Reject H_0, if $\mid T \mid > c$
Given H_0, $T \sim \text{Normal}(0, \sigma_T^2)$, where $\sigma_T^2 = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(\mid T \mid > c \mid \mu_1 = \mu_2)\\ =& 2F_Z\left(\frac{-c}{\sigma_T}\right)\end{aligned}$$

- The null and alternative hypothesis are:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 > \mu_2$$

Test Statistic: $T = \overline{X} - \overline{Y}$
Test: Reject H_0, if $\overline{X} - \overline{Y} > c$
Given H_0, $T \sim \text{Normal}(0, \sigma_T^2)$, where $\sigma_T^2 = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(\overline{X} - \overline{Y} > c \mid \mu_1 = \mu_2)\\ =& 1 - F_Z\left(\frac{c}{\sigma_T}\right)\end{aligned}$$

- The null and alternative hypothesis are:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 < \mu_2$$

Test Statistic: $T = \overline{X} - \overline{Y}$
Test: Reject H_0, if $\overline{Y} - \overline{X} > c$
Given H_0, $T \sim \text{Normal}(0, \sigma_T^2)$, where $\sigma_T^2 = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(\overline{Y} - \overline{X} > c \mid \mu_1 = \mu_2)\\ =& 1 - F_Z\left(\frac{c}{\sigma_T}\right)\end{aligned}$$

5. **Two samples F-test (known variances)**

Let $X_1, \ldots, X_{n_1} \sim$ iid Normal(μ_1, σ_1^2)
and $Y_1, \ldots, Y_{n_2} \sim$ iid Normal(μ_2, σ_2^2)
Following are the three different possibilities:

- The null and alternative hypothesis are:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 > \sigma_2$$

  Test Statistic: $T = \dfrac{S_1^2}{S_2^2}$
  Test: Reject H_0, if $T > 1 + c$
  Given H_0, $T \sim F(n_1 - 1, n_2 - 1)$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(T > 1 + c \mid \sigma_1 = \sigma_2)\\ =& 1 - F_{F(n_1-1,n_2-1)}(1 + c)\end{aligned}$$

- The null and alternative hypothesis are:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 < \sigma_2$$

  Test Statistic: $T = \dfrac{S_1^2}{S_2^2}$
  Test: Reject H_0, if $T < 1 - c$
  Given H_0, $T \sim F(n_1 - 1, n_2 - 1)$

$$\begin{aligned}\alpha =& P(\text{reject } H_0 \mid H_0 \text{ is true})\\ =& P(T < 1 - c \mid \sigma_1 = \sigma_2)\\ =& F_{F(n_1-1,n_2-1)}(1 - c)\end{aligned}$$

- The null and alternative hypothesis are:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 \neq \sigma_2$$

  Test Statistic: $T = \dfrac{S_1^2}{S_2^2}$
  Test: Reject H_0, if $T > 1 + c_R$ or $T < 1 - c_L$
  Given H_0, $T \sim F(n_1 - 1, n_2 - 1)$

$$\frac{\alpha}{2} = P(T > 1 + c_R \mid H_0) = P(T < 1 - c_L \mid H_0)$$

6. **Likelihood Ratio test**:
For simple null and alternative hypothesis, Likelihood ratio test is enough.

$$X_1, \ldots, X_n \sim P$$

Consider the simple null and alternative hypothesis:

$$H_0 : P = f_X$$

$$H_A : P = g_X$$

Likelihood ratio: $L(X_1, \ldots, X_n) = \dfrac{\prod_{i=1}^{n} g_X(X_i)}{\prod_{i=1}^{n} f_X(X_i)}$

Likelihood ratio test: Reject H_0, if $T = L(X_1, \ldots, X_n) > c$

7. **χ^2-test for goodness of fit:**
H_0 : Samples are i.i.d X, $\quad H_A$: Samples are not i.i.d X

Test statistic: $T = \displaystyle\sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{k-1}$

Test: Reject H_0 if $T > c$.
Significance level: $\alpha = P(T > c \mid H_0) \approx 1 - F_{\chi^2_{k-1}}(c)$

Note: In case of continuous distribution, convert continuous to discrete by binning.

8. **Test for independence:**
H_0 : Joint PMF is product of marginals, H_A : Joint PMF is not product of marginals

Test statistic: $T = \displaystyle\sum_{i,j} \frac{(y_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{dof}$

where dof = (number of rows-1) $\times$ (number of columns-1)
y_{ij} = product of marginals for (i, j)
np_{ij} = expected, if independent

Test: Reject H_0 if $T > c$.