Review of probability theory (Continued)

Variance:

\[ \text{var}(X) = E \left( (X - E(X))^2 \right) = E \left( X^2 - 2XE(X) + (E(X))^2 \right) \]

\[ = E(X^2) - 2E(X)E(X) + (E(X))^2 = E(X^2) - (E(X))^2 \]

We obtain:

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

Standard deviation:

\[
\text{std}(X) = \sqrt{\text{var}(X)}
\]

Properties of \(E(X)\) and \(\text{var}(X)\)

i) \(E(aX + bY) = aE(X) + bE(Y)\)

This is valid for all \(X\) and \(Y\).

In particular, \(X\) and \(Y\) do not need to be independent.

ii) If \(X\) and \(Y\) are independent, then we have

\[ E(XY) = E(X)E(Y) \]

Proof:

\[
E(XY) = \int xy \rho_{(X,Y)}(x,y) dx dy = \int xy \rho_x(x) \rho_y(y) dx dy
\]

\[ = \left( \int x \rho_x(x) dx \right) \left( \int y \rho_y(y) dy \right) = E(X)E(Y) \]

Caution: \(E(XY) = E(X)E(Y)\) does not imply that \(X\) and \(Y\) are independent.

Example:
\((X,Y) = \begin{cases} (0,1), & \Pr = 0.25 \\ (0,-1), & \Pr = 0.25 \\ (1,0), & \Pr = 0.25 \\ (-1,0), & \Pr = 0.25 \end{cases}\)

\[E(X) = 0, \quad E(Y) = 0, \quad E(XY) = 0\]

But \(X^2 + Y^2 = 1\)

iii) If \(X\) and \(Y\) are independent, then we have

\[\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)\]

**Proof:**

\[\text{var}(X + Y) = E((X + Y)^2) - (E(X + Y))^2 = \cdots\]

Complete the proof in your homework.

**Examples of distributions:**

1) Bernoulli distribution

\[X = \begin{cases} 1, & \Pr = p \\ 0, & \Pr = 1 - p \end{cases}\]

**Example:** Flip a coin

1: head, success

0: tail, failure

Range = \{0, 1\}

Expected value and variance:

\[E(X) = p, \quad E(X^2) = p\]

\[\text{Var}(X) = E(X^2) - (E(X))^2 = p(1-p)\]

2) Binomial distribution

\[N = \sum_{i=1}^{n} X_i\]

PMF (probability mass function):
\[ \Pr(N = k) = C(n, k) p^k (1 - p)^{n-k} \]

**Example:**  # of heads in \( n \) flips

Range = \{0, 1, 2, ..., \( n \)\}

Expected value and variance:

\[ E(N) = E(X_1 + X_2 + \cdots + X_n) = np \]

\[ \text{var}(N) = \text{var}(X_1 + X_2 + \cdots + X_n) = np(1 - p) \]

3) **Exponential distribution**

**PDF (probability density function):**

\[ \rho_{\tau}(t) = \begin{cases} \lambda \exp(-\lambda t), & t \geq 0 \\ 0, & t < 0 \end{cases} \]

**Example:** (Escape problem)

\( T \) = time until escape from a deep potential well by thermal fluctuations

**Mathematical definition:**

\( T \) = time until occurrence of an event in a memoryless system

We derive the PDF based on the “memoryless” property.

“memoryless” means

\[ \Pr((T - t_0) \leq t \mid T > t_0) = \Pr(T \leq t) \]

Consider the complementary cumulative distribution function (CCDF)

\[ G(t) = \Pr(T > t) = \int_{t}^{\infty} \rho(t') dt' \]

\[ G(0) = \Pr(T > 0) = 1 \]

We re-write the memoryless property in terms of \( G(t) \).

\[ \frac{\Pr((T - t_0) \leq t \text{ AND } T > t_0)}{\Pr(T > t_0)} = \Pr(T \leq t) \]

\[ \implies \Pr(t_0 < T \leq t_0 + t) = \Pr(T \leq t) \Pr(T > t_0) \]

\[ \implies G(t_0) - G(t_0 + t) = (1 - G(t)) G(t_0) \]

Replace \( t \) with \( \Delta t \), divide by \( \Delta t \), and take the limit as \( \Delta t \to 0 \), we get
\[
\frac{G(t_0) - G(t_0 + \Delta t)}{\Delta t} = \frac{G(0) - G(\Delta t)}{\Delta t} G(t_0)
\]

\[
\Rightarrow G'(t_0) = \frac{G'(0)}{-\lambda} G(t_0)
\]

We obtain an initial value problem (IVP)

\[
\begin{cases}
G'(t_0) = -\lambda G(t_0) \\
G(0) = 1
\end{cases}
\]

The solution is \( G(t) = \exp(-\lambda t) \).

Differentiate \( G(t) \), we obtain

\[
\rho(t) = -\frac{d}{dt} G(t) = \begin{cases}
\lambda \exp(-\lambda t), & t \geq 0 \\
0, & t < 0
\end{cases}
\]

Expected value and variance:

\[
E(T) = \frac{1}{\lambda}, \quad \text{var}(T) = \frac{1}{\lambda^2}
\]

CDF:

\[
F_T(t) = \Pr(T \leq t) = 1 - \exp(-\lambda t) \quad \text{for } t \geq 0
\]

4) Normal distribution

PDF:

\[
\rho_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\frac{-(x-\mu)^2}{2\sigma^2}\right)
\]

Notation: \( X \sim N(\mu, \sigma^2) \)

Example: (Central Limit Theorem)

Suppose \( \{X_1, X_2, ..., X_M\} \) are i.i.d.

(independently and identically distributed).

When \( M \) is large, approximately \( X = \sum_{j=1}^{M} X_j \) has a normal distribution.

Range = \((-\infty, +\infty)\)
Expected value and variance:
\[
E(X) = \int x p(x) dx = \mu
\]
\[
\text{var}(X) = E((X - \mu)^2) = \int (x - \mu)^2 p(x) dx = \sigma^2
\]

CDF of normal distribution:
\[
F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x' - \mu)^2}{2\sigma^2} \right) dx'
\]

Change of variables: 
\[
s = \frac{x' - \mu}{\sqrt{2\sigma^2}}, \quad dx' = \sqrt{2\sigma^2} ds
\]

\[
F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{\pi}} \exp(-s^2) ds = \frac{1}{2} + \int_{0}^{x} \frac{1}{\sqrt{\pi}} \exp(-s^2) ds
\]

The error function:
\[
\text{erf}(z) \equiv \frac{1}{\sqrt{\pi}} \int_{-z}^{z} \exp(-s^2) ds = \frac{2}{\sqrt{\pi}} \int_{0}^{z} \exp(-s^2) ds
\]

With the error function, we write the CDF as
\[
F_X(x) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{x - \mu}{\sqrt{2\sigma^2}} \right) \right)
\]

Example:
\[
\Pr(X \leq \mu + \eta \sigma) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{\mu + \eta \sigma - \mu}{\sqrt{2\sigma^2}} \right) \right) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{\eta}{\sqrt{2}} \right) \right)
\]

Properties of \text{erf}(z):
\[
i) \quad \text{erf}(0) = 0
\]
\[
ii) \quad \text{erf}(+\infty) = 1
\]
\[
iii) \quad \text{erf}(-z) = \text{erf}(z)
\]

Shape of PDF:
\[
\text{(Draw the bell-shaped PDF, showing the center } \mu, \text{ and show the interval } [\mu - \eta \sigma, \mu + \eta \sigma] \text{.)}
\]
We like to find \( \eta \) such that
\[
\Pr(|X-\mu| \leq \eta \sigma) = 0.95 \quad (95\%)
\]
We express this probability in terms of CDF.
\[
\Pr(|X-\mu| \leq \eta \sigma) = \Pr(\mu - \eta \sigma \leq X \leq \mu + \eta \sigma)
\]
\[
= F_X(\mu + \eta \sigma) - F_X(\mu - \eta \sigma) = \cdots = \text{erf}\left(\frac{\eta}{\sqrt{2}}\right)
\]
We set \( \text{erf}\left(\frac{\eta}{\sqrt{2}}\right) = 0.95 \)
\[
\Rightarrow \eta = \text{erfinv}(0.95)\sqrt{2} = 1.96
\]
We obtain
\[
\Pr(|X-\mu| \leq 1.96\sigma) = 95\%
\]
Similarly, we can obtain
\[
\Pr(|X-\mu| \leq 2.5758\sigma) = 99\%
\]

Confidence interval:

Suppose we are given a data set of \( n \) independent samples of \( X \sim N(\mu, \sigma^2) \).
\( \{X_j, j = 1, 2, \ldots, n\} \)

Question: How to estimate \( \mu \) from data?
\[
\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j
\]

Question: How to estimate the uncertainty in \( \hat{\mu} \)?
\( \hat{\mu} \) is a random variable.
\[
E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{j=1}^{n} X_j\right) = \frac{1}{n} E\left(X_1 + \cdots + X_n\right) = \mu
\]
\[
\text{var}(\hat{\mu}) = \text{var}\left(\frac{1}{n} \sum_{j=1}^{n} X_j\right) = \frac{1}{n^2} \text{var}\left(X_1 + \cdots + X_n\right) = \frac{\sigma^2}{n}
\]
(Here we used the independence of \( \{X_i\} \))

Theorem:
This theorem will be proved later in the discussion of characteristic functions. It follows that \( \hat{\mu} \) is normal.

\[
\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)
\]

The interval containing 95% probability of \( \hat{\mu} \) is described by

\[
\Pr\left( |\hat{\mu} - \mu| \leq 1.96 \frac{\sigma}{\sqrt{n}} \right) = 95\%
\]

**Case 1:** Suppose we know the value of \( \sigma \).

\[
\left( \hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]

is called the 95% confidence interval.

**Example:**

We are given a data set of 100 independent samples of \( X \sim N(\mu, \sigma^2) \).

\{3.0811, 0.7589, 1.9611, \ldots \}

We are given \( \sigma = 1.3 \)

\( \mu \) is estimated as

\[
\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j = 0.475
\]

\[
1.96 \frac{\sigma}{\sqrt{n}} = 0.2548
\]

The 95% confidence interval is \( (0.2202, 0.7298) \)

**Question:** What is the meaning of this confidence interval?

The data set is given, fixed.

\( \mu \) is fixed, although unknown.

What is the meaning of \( (0.2202, 0.7298) \)?

**Two key components in interpreting the confidence interval:**

i) The confidence interval is an algorithm/function that maps data to interval

\[
\left( \hat{\mu}_L \left( \{X_j\} \right), \hat{\mu}_U \left( \{X_j\} \right) \right)
\]

ii) The framework of repeated experiments.

Draw a data set of \( n \) independent samples of \( X \sim N(\mu^{(\text{True})}, \sigma^2) \).
Repeat this $M$ times ($M$ is large). The meaning of confidence interval is

$$\Pr \left( \frac{\hat{\mu}_j(\{X_j\})}{\text{Random variable}} < \frac{\mu^{(\text{True})}}{\text{Fixed}} < \frac{\hat{\mu}_n(\{X_j\})}{\text{Random variable}} \right) = 0.95$$

In other words, suppose we go over $M$ data sets and estimate the confidence interval for each data set. For 95% of data sets, the confidence interval contains $\mu^{(\text{True})}$.

A similar example: I have an algorithm for estimating an interval for a dog’s age. Suppose we go over $M$ dogs. For 95% of dogs, the dog’s true age falls in the estimated interval.

In summary, the two key components are i) an algorithm, ii) repeated experiments.

**Case 2:** $\sigma$ is unknown

Recall the definition of standard deviation.

$$\sigma = \sqrt{\text{var}(X)} = \sqrt{E\left( (X - \mu)^2 \right)}$$

We can estimate $\sigma$ as

$$\hat{\sigma} = \frac{1}{n} \sum_{j=1}^{n} (X_j - \hat{\mu})^2, \quad \hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

An approximate 95% confidence interval is

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

Correspondingly, the one calculated using the exact value of $\sigma$ (case 1 above) is called the exact 95% confidence interval

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