Review of Basic Probability-II

Based on “Introduction to Mathematical Statistics” by Hogg, McKenna and Craig

CS698C: Sketching and Sampling for Big Data Analysis

IIT Kanpur

Outline

Bernoulli Experiment

- *Bernoulli experiment* is a random experiment which has exactly two outcomes, classified as for instance, success or failure (or, as life/death, male/female, defective/non-defective).
- The random variable X is defined as, for example,

$$X(\text{success}) = 1 \quad \text{and} \quad X(\text{failure}) = 0 \ .$$

- Probabilities are associated with the two outcomes:

$$\mathrm{P}[X = 1] = p \quad \text{and} \quad \mathrm{P}[X = 0] = 1 - p \ .$$

  Equivalently this is the pmf of X.
- The expectation of X is:

$$\mathrm{E}[X] = 0 \cdot (1 - p) + 1(p) = p \ .$$

- The expectation of X^2 is:

$$\mathrm{E}\left[X^2\right] = 0^2 \cdot (1 - p) + 1^2(p) = p \ .$$

Bernoulli Distribution

- So variance of X is

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]\right)^2 = p - p^2 = p(1-p) \ .$$

- Mgf is:

$$\mathrm{E}\left[e^{tX}\right] = e^{t\cdot 0}(1-p) + e^{t\cdot 1}p = 1 - p + pe^t \ .$$

  This is defined for all t, $-\infty < t < \infty$.

Binomial Distribution

- Consider an experiment with a sequence of n Bernoulli trials. Let X_i denote the Bernoulli random variable corresponding to the ith trial.
- The outcome observed is an n-tuple of 0s and 1s.
- We are often interested in the total number of successes and not in the order of occurrence.
- Now let X denote the number of observed successes in the n Bernoulli trials, then, the space for X is $0, 1, 2, \ldots, n$.
- If k successes occur, then, the number of ways of selecting the positions of the k successes in the n trials is

$$\binom{n}{k}$$

- Since trials are independent, the probability of the outcome of a given sequence with k successes and $n - k$ failures is, by independence of trials exactly

$$p^k (1-p)^{n-k} \quad .$$

Binomial Distribution

- The pmf of X is

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n \ .$$

- This is a valid pmf, since, $p(k) \geq 0$, for each $k = 0, 1, \ldots, n$ and

$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$

  by the binomial theorem $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$.

- The binomial distribution with parameters n and p is often denoted as $B(n, p)$ or as $b(n, p)$.

Binomial Distribution

- Let X be a random variable which is the sum of n independent Bernoulli random variables, each with probability of success p.
- Denoting the random variable corresponding to the ith Bernoulli trial as X_i, we have
$$X = X_1 + X_2 + \cdots + X_n \quad .$$
- We have $\mathrm{E}[X_i] = p$. By linearity of expectation,
$$\mathrm{E}[X] = \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = p + \cdots + p = np \quad .$$

Binomial Distribution: Variance

- Likewise, variance is calculated.

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n}(X_i - p)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n}(X_i - p)^2 + 2\sum_{1\le i<j\le n}(X_i - p)(X_j - p)\right] \\
&= \sum_{i=1}^{n}\mathrm{E}\left[(X_i - p)^2\right] + 2\sum_{1\le i<j\le n}\mathrm{E}\left[(X_i - p)(X_j - p)\right] \\
&= \sum_{i=1}^{n}\mathrm{Var}[X_i] + 2\sum_{1\le i<j\le n}\mathrm{E}[X_i - p]\,\mathrm{E}[X_j - q] \\
&= np(1-p) + 2\sum_{1\le i<j\le n} 0\cdot 0 \\
&= np(1-p)\ .
\end{aligned}
$$

Binomial Distribution: Mgf

- Writing the binomially distributed $B(n, p)$ variable X as the sum of the successes in n independent Bernoulli trials, where, $X_i = 1$ if the ith Bernoulli trial is 1, and $X_i = 0$ otherwise, we have,

$$X = X_1 + \cdots + X_n \quad .$$

- Then, by independence and properties of mgf,

$$\begin{aligned}
\mathrm{E}\left[e^{tX}\right] &= \mathrm{E}\left[e^{t(X_1+\cdots+X_n)}\right] \\
&= \mathrm{E}\left[e^{tX_1}\right]\mathrm{E}\left[e^{tX_2}\right]\cdots\mathrm{E}\left[e^{tX_n}\right] \\
&= (1-p+pe^t)(1-p+pe^t)\cdots(1-p+pe^t) \\
&= (1-p+pe^t)^n \quad .
\end{aligned}$$

- mgf is defined for all values of t, $-\infty < t < \infty$.

Weak law of large numbers

- Let X be the number of successes in n independent Bernoulli trials, each with probability of success p.
- Define the variable $Y = X/n$.
- Then,

$$\mathrm{E}[Y] = \frac{1}{n}\mathrm{E}[X] = \frac{(np)}{n} = p$$

  and

$$\mathrm{Var}[Y] = \frac{1}{n^2}\mathrm{Var}[X] = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n} \quad .$$

- By Chebychev's inequality, for any fixed $\epsilon > 0$,

$$\mathrm{P}\left[\left|\frac{X}{n} - p\right| \geq \epsilon\right] = \mathrm{P}[|Y - p| \geq \epsilon] \leq \frac{\mathrm{Var}[Y]}{\epsilon^2} = \frac{p(1-p)}{\epsilon^2 n}$$

- **(Weak law of large numbers)** Therefore, as $n \to \infty$,

$$\lim_{n\to\infty} \mathrm{P}\left[\left|\frac{X}{n} - p\right| \geq \epsilon\right] = 0$$

Example: Median of n random variables

- Let $X_1, X_2, \ldots, X_n$ be iid random variables.
- Let Y be the middle value or median of $X_1, \ldots, X_n$.
- Find the cdf of Y. Denote it as $F_Y(y)$.
- For any fixed y, define the "success" event as $X_i \leq y$ and the "failure" event as $X_i > y$, $i = 1, 2 \ldots, n$.
- That is, define W_i to be Bernoulli variable

$$W_i = \begin{cases} 1 & \text{if } X_i \leq y \\ 0 & \text{otherwise.} \end{cases}$$

- The W_i's are independent and identically distributed with probability of success

$$\mathrm{P}\left[W_i = 1\right] = p = F_X(y) \quad .$$

Median Distribution

- For any fixed y, $Y \leq y$ holds if at least half of the X_i's satisfy $X_i \leq y$.
- Equivalently, if we define $W = W_1 + W_2 + \cdots + W_n$, then, the two events are equivalent:
$$Y \leq y \qquad \equiv W \geq n/2$$
- For simplicity, assume that n is even so that $n/2$ is integral. So, we have,
$$\begin{aligned} F_Y(y) &= \mathrm{P}[Y \leq y] \\ &= \mathrm{P}[W \geq n/2] \\ &= \sum_{k=n/2}^{n} \binom{n}{k} (F_X(y))^k (1 - F_X(y))^{n-k} \quad . \end{aligned}$$

Example contd.

- The pdf is given by

$$
\begin{aligned}
f_Y(y) &= F_Y'(y) \\
&= \sum_{k=n/2}^{n-1} \binom{n}{k} \Big[k(F_X(y))^{k-1}(1 - F_X(y))^{n-k} f_X(y) \\
&\qquad -(F_X(y))^k (1 - F_X(y))^{n-k-1} f_X(y)\Big] \\
&\qquad + n(F_X(y))^{k-1} f_X(y)
\end{aligned}
$$

Geometric Distribution

- Consider a sequence of coin tosses, independently and with constant probability of heads is p.
- Let the random variable X be the number of tosses before the first heads appears (including the toss resulting in first heads).
- Then, X takes values $1, 2, 3, \ldots,$.
- What is $P(X = k)$.
- $X = k$ iff the first $k - 1$ tosses are a failure and the kth toss is a success. By independence, this probability is

  $$\mathrm{P}[X = k] = (1 - p)^{k-1} p, \quad k = 1, 2, \ldots \quad .$$

  (Show that this is a probability distribution).
- This is called a *geometric distribution*.

A variant of Geometric Distribution

- Consider the same experiment as before. Let Y denote the number of coin tosses before the first heads appears, but not including the first heads toss.
- Then, Y takes values from $0, 1, 2, \ldots$, and by independence,

$$\mathrm{P}[Y = k] = (1 - p)^k p, \qquad k = 0, 1, 2, \ldots \quad .$$

- Note that $\mathrm{P}[Y = k] = \mathrm{P}[X = k + 1]$, from the previous example.

Extension: Negative Binomial Distribution

- Consider the same experiment as before. So the outcome is a sequence of heads and tails.
- Define the event Y to be the number of tail coin tosses before the rth occurrence of heads. (not including the rth occurrence).
- So Y takes values from $0, 1, 2, \ldots,$.
- What is $\mathrm{P}[Y = y]$?
- Out of the first $y + r - 1$ coin tosses, there are exactly $r - 1$ heads, and the remaining y tails. The $(y + r)$th coin toss is a heads.
- So $Y = y$ iff there are $r - 1$ heads in the first $y + r - 1$ coin tosses. Additionally the $(y + r)$th coin toss is a heads.
- The number of heads in the first $y + r - 1$ coin tosses is a binomial distribution $B(y + r - 1, p)$. Hence,

$$\begin{aligned}\mathrm{P}[Y = y] = p_Y(y) &= \binom{y+r-1}{r-1} p^{r-1}(1-p)^y \cdot p \\ &= \binom{y+r-1}{r-1} p^r (1-p)^y, \quad y = 0, 1, 2, \ldots,\end{aligned}$$

Negative Binomial Distribution

- It is called the negative binomial distribution because of the identity,

$$(1-q)^{-r} = \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} q^y, \quad 0 < q < 1 \ .$$

- Here $q = 1 - p$.
- The pmf $p_Y(y)$ satisfies

$$\begin{aligned}
&\sum_{y=0}^{\infty} p_Y(y) \\
&= \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} p^r (1-p)^y \\
&= p^r \left(1 - (1-p)\right)^{-r} && \text{negative binomial theorem identity} \\
&= 1 \ .
\end{aligned}$$

Multinomial Distribution

- Generalizes binomial distribution.
- A random experiment is repeatedly performed n times independently.
- The outcome of each experiment is exactly one of k possibilities, say $C_1, C_2, \ldots, C_k$. These are mutually exclusive and exhaustive.
- Let p_1 be the probability that C_1 occurred, p_2 be the probability that C_2 occured, and so on, till p_k.
- So $p_1 + p_2 + \cdots + p_k = 1$.
- The random experiment is repeated n times. Let X_i be the number of outcomes where C_i occurred, $i = 1, 2, \ldots, k-1$.
- Note that X_k is not explicitly defined, since, $X_k = n - X_1 - \cdots - X_{k-1}$.

Pmf

- Let $x_1, x_2, \ldots, x_{k-1}$ be non-negative integers so that $x_1 + x_2 + \cdots + x_n \leq n$.
- The probability that out of the n experiments, C_1 occurred x_1 times, C_2 occurred x_2 times, $\cdots$, C_{k-1} occurred x_{k-1} times is (by independence) given as follows. In this case C_k occurs exactly $x_k = n - x_1 - x_2 - \cdots - x_{k-1}$ times.

$$
\begin{aligned}
&p(x_1, x_2, \ldots, x_{k-1}) \\
&\qquad = \binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-2}}{x_{k-1}} p_1^{x_1} p_2^{x_2} \cdots p_{k-1}^{x_{k-1}} p_k^{x_k} \quad .
\end{aligned}
$$

- Simplifying the expression

$$
\begin{aligned}
&\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-2}}{x_{k-1}} \\
&= \frac{n!}{x_1!(n-x_1)!}\frac{(n-x_1)!}{x_2!(n-x_1-x_2)!}\cdots\frac{(n-x_1-\cdots-x_{k-2})!}{x_{k-1}!(n-x_1-\cdots-x_{k-1})!} \\
&= \frac{n!}{x_1!x_2!\cdots x_{k-1}!(n-x_1-\cdots-x_{k-1})!}
\end{aligned}
$$

Pmf of Multinomial distribution

- Noting that $x_k = n - x_1 - x_2 - \cdots - x_{k-1}$, This gives

$$p(x_1, x_2, \ldots, x_{k-1}) = \frac{n!}{x_1!x_2!\cdots x_{k-1}!x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

  for $0 < x_1, x_2, \ldots, x_{k-1} \leq 1$ and $x_1 + \cdots + x_{k-1} \leq 1$.
- The probability of each outcome with $x_1, \ldots, x_{k-1}$ is $p_1^{x_1} \cdots p_k^{x_k}$.
- $\binom{n}{x_1, x_2, \ldots, x_k}$ is the number of outcomes with $x_1, \ldots, x_{k-1}$.
- Multinomial theorem:

$$(a_1 + a_2 + \cdots + a_k)^n = \sum_{\substack{x_1 \geq 0, \cdots, x_{k-1} \geq 0 \\ x_1 + \ldots + x_{k-1} \leq n}} \frac{n!}{x_1!x_2!\ldots x_k!} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}$$

Pmf of multinomial distribution

- Therefore,

$$
\begin{aligned}
&\sum_{\substack{x_1 \geq 0, \cdots, x_{k-1} \geq 0 \\ x_1 + \ldots + x_{k-1} \leq n}} p(x_1, x_2, \ldots, x_{k-1}) \\
&= \sum_{\substack{x_1 \geq 0, \cdots, x_{k-1} \geq 0 \\ x_1 + \ldots + x_{k-1} \leq n}} \frac{n!}{x_1! x_2! \cdots x_{k-1}! x_k!} p_1^{x_1} \cdots p_k^{x_k} \\
&= (p_1 + p_2 + \cdots + p_n)^n \\
&= 1
\end{aligned}
$$

Marginal pdfs for multinomial distribution

- Given a multinomial distribution for a random vector $(X_1, X_2, \ldots, X_{k-1})$ with parameters n, k and $p_1, p_2, \ldots, p_n$.
- Find the marginal distributions of X_i, $i = 1, \ldots, k$.

$$
\begin{aligned}
p_{X_1}(x_1) &= \binom{n}{x_1} p_1^{x_1} \sum_{x_2=0}^{n-x_1} \sum_{x_3=0}^{n-x_1-x_2} \cdots \sum_{x_{k-1}=0}^{n-x_1-x_2-\cdots-x_{k-2}} \frac{(n-x_1)!}{x_2! x_3! \cdots x_{k-1}! x_k!} p_2^{x_2} \cdots p_k^{x_k} \\
&= \sum_{\substack{x_2 \geq 0, \ldots, x_{k-1} \geq 0 \\ x_2 + \cdots + x_{k-1} \leq n - x_1}} \binom{n - x_1}{x_2 \; x_3 \; \ldots \; x_{k-1} \, x_k} p_2^{x_2} \cdots p_k^{x_k} \\
&= \binom{n}{x_1} p_1^{x_1} (p_2 + p_3 + \cdots + p_k)^{n-x_1} \\
&= \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}
\end{aligned}
$$

- So marginal pdf of X_1 is a binomial distribution with parameters n and p_1; $B(n; p_1)$. Similarly marginal pdf of X_i is a binomial distribution-$B(n, p_i)$.

Multinomial Distribution: Conditional Distribution

- Let us consider the conditional pdf of $X_1 \mid X_2 = x_2$. We wish to find $p_{X_1}(x_1 \mid x_2)$.
- Consider the space of all outcome sequences where $X_2 = x_2$. There are $\binom{n}{x_2}$ such sequences and its probability is $p_{X_2}(x_2) = \binom{n}{x_2} p_2^{x_2}(1-p_2)^{n-x_2}$.
- For any sequence where the positions of the occurrence of C_2 are now fixed, the x_1 positions where C_1 occurs can happen in $\binom{n-x_2}{x_1}$ ways. Its probability is $\binom{n-x_2}{x_1} p_1^{x_1}(1-p_1-p_2)^{n-x_1-x_2}$, since in the remaining $n-x_2-x_1$ positions, neither C_1 nor C_2 can occur.

Conditional Distribution

- The conditional distribution is

$$p_{X_1|x_2}(x_1 \mid x_2) = \frac{\binom{n}{x_2} p_2^{x_2} \binom{n-x_2}{x_1} p_1^{x_1} (1-p_1-p_2)^{n-x_2-x_1}}{\binom{n}{x_2} p_2^{x_2} (1-p_2)^{n-x_2}}, \quad 0 \le x_1 \le n - x_2$$

$$= \binom{n-x_2}{x_1} \left(\frac{p_1^{x_1}}{(1-p_2)^{x_1}} \right) \left(\frac{(1-p_1-p_2)^{n-x_2-x_1}}{(1-p_2)^{n-x_2-x_1}} \right)$$

$$= \binom{n-x_2}{x_1} \left(\frac{p_1}{1-p_2} \right)^{x_1} \left(1 - \frac{p_1}{1-p_2} \right)^{n-x_2-x_1}$$

- $X_1 \mid X_2 = x_2$ is distributed as a binomial distribution, namely, $B(n - x_2, \frac{p_1}{1-p_2})$.
- Hence,

$$\mathrm{E}[X_1 \mid x_2] = (n - x_2) \left(\frac{p_1}{1-p_2} \right)$$

and is a linear function of x_2.

Hypergeometric Distribution

- Hypergeometric distribution is a random sampling without replacement.
- Suppose there is a bin containing N balls, out of which r balls are red and $g = N - r$ balls are green.
- Suppose we choose m balls at random ***without replacement***. Assume $m \leq r, m \leq N - r$. What is the probability that there are x red balls in the sample.
- Note: The red balls may be viewed as defective items and green balls as normal items. We wish to find the probability of finding x defective items in a sample of m balls, chosen at random but *without replacement*.

Hypergeometric Distribution

- The number of ways of choosing m balls from N balls is $\binom{N}{m}$. Each of these choices of m ball groups has the same probability which is $1/\binom{N}{m}$.
- The number of ways of choosing a sample containing x red balls and $m - x$ green balls is $\binom{r}{x}\binom{N-r}{m-x}$.
- Hence,

$$p(x) = \frac{\binom{r}{x}\binom{N-r}{m-x}}{\binom{N}{m}}, \quad x = 0, 1, \ldots, m.$$

- X is a hypergeometric distribution with parameters N, r and m.

Hypergeometric Distribution

- Using the fact that $\binom{n}{m} = \frac{n}{m}\binom{n-1}{m-1}$, for $1 \leq m \leq n$,

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{x=0}^{m} \frac{x\binom{r}{x}\binom{N-r}{m-x}}{\binom{N}{m}} \\
&= \sum_{x=1}^{r} \frac{x(r/x)\binom{r-1}{x-1}\binom{(N-1)-(r-1)}{(m-1)-(x-1)}}{(N/m)\binom{N-1}{m-1}} \\
&= \frac{mr}{N}\left[\sum_{x-1=0}^{r-1} \frac{\binom{r-1}{x-1}\binom{(N-1)-(r-1)}{(m-1)-(x-1)}}{\binom{N-1}{m-1}}\right] \\
&= \frac{r}{N}m
\end{aligned}
$$

  since, the expression in the square brackets is 1 as it is the pmf of hypergeometric distribution with parameters $N-1, r-1$ and $m-1$.

- The expectation for sampling with replacement (Binomial $(m, p = r/N)$) and sampling without replacement (Hypergeometric above) are the same.

Poisson Distribution

- The Poisson distribution has a pmf with parameter m.

$$p(x) = \frac{m^x e^{-m}}{x!}, \quad x = 0, 1, 2, \ldots, \ .$$

- It is pmf since,

$$\sum_{x=0}^{\infty} \frac{m^x e^{-m}}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{m^x}{x!} = e^{-m} \cdot e^m = 1 \ .$$

- Let X be a random variable with a Poisson distribution. Then, $\mathrm{E}[X] = m$.

$$\mathrm{E}[X] = \sum_{x=0}^{\infty} \frac{x m^x e^{-m}}{x!} = m e^{-m} \sum_{x=1}^{infty} \frac{m^{x-1}}{(x-1)!} = m e^{-m} e^m = m \ .$$

- So, the mean μ is said to denote the parameter of the Poisson distribution.

Poisson Process

- Consider a modeling of the number of arrivals x in an interval of time w. It could be the number of calls in a telecom switch, or packets in a network switch, etc..
- The Poisson postulates for this modeling are as follows.
- Let $g(x, w)$ denote the probability of x arrivals in an interval of time w. In the following $h > 0$ is small, and $\lambda > 0$ is a positive constant. The symbol $o(h)$ represents any function such that $\lim_{h\to 0} \frac{o(h)/h}{=} 0$.
    1. $g(1, h) = \lambda h + o(h)$.
    2. $\sum_{x=2}^{\infty} g(x, h) = o(h)$.
    3. The number of arrivals in non-overlapping intervals are independent.
- Postulates 1 and 3 state that effectively the probability of one arrival in a short interval is approximately proportional to the length of the interval, and is independent of arrivals/non-arrival in other non-overlapping intervals.
- Postulate 2 states that the probability of two or more arrivals within a short interval h tends to 0 in the limit $h \to 0$.

Poisson process

- It can be shown that the Poisson postulates give the following solution

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}, x = 1, 2, 3, \ldots, \ .$$

- Note that for each fixed w, $p_w(x) = g(x, w)$ is the pmf of the Poisson distribution with parameter λw.
- Thus, the number of arrivals X in an interval of length w has a Poisson distribution with parameter $\mu = \lambda w$.
- The function $g(x, w)$ is defined for real $w \geq 0$ and for $x = 1, 2, 3, \ldots,$.

Poisson Distribution

- Let X have a Poisson distribution with parameter μ. pmf is $p(x) = \frac{\mu^x e^{-\mu}}{x!}$, $x = 1, 2, 3, \ldots$ and zero elsewhere.
- $\mathrm{E}[X] = \mu$.
- $\mathrm{Var}[X] = \mu$.

$$\mathrm{E}[X(X-1)] = \sum_{x=0}^{\infty} \frac{x(x-1)\mu^x e^{-\mu}}{x!} = e^{-\mu}\mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} = \mu^2 e^{-\mu} e^{\mu} = \mu^2 \quad .$$

- Hence,

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mu^2 = \mathrm{E}[X(X-1) + X] - \mu^2 \\ &= \mathrm{E}[X(X-1)] + \mu - \mu^2 = \mu^2 + \mu - \mu^2 = \mu \quad .\end{aligned}$$

Outline

Poisson, Gamma and χ^2 distributions

Mgf of Poisson Distribution

- Mgf of Poisson distribution with parameter μ is

$$M(t) = \mathrm{E}\left[e^{tX}\right] = e^{\mu(e^t-1)}$$

$$\mathrm{E}\left[e^{tX}\right] = \sum_{x=0}^{\infty} \frac{e^{tx}\mu^x e^{-\mu}}{x!} = e^{-\mu}\sum_{x=0}^{\infty}\frac{(e^t\mu)^x}{x!} = e^{-\mu}e^{e^t\mu} = e^{\mu(e^t-1)} \quad .$$

- A nice property of Poisson distributions is the **additive property**.
- Let $X_1, \ldots, X_n$ be independent random variables such that X_i has a Poisson distribution with parameter μ_i. Then, $Y = X_1 + X_2 + \cdots + X_n$ has a Poisson distribution with parameter $\mu_1 + \cdots + \mu_n$.
- Pf: Follows from the uniqueness of the mgfs.

$$\begin{aligned}\mathrm{E}\left[e^{t(X_1+\cdots+X_n)}\right] &= \mathrm{E}\left[e^{tX_1}\right]\mathrm{E}\left[e^{tX_2}\right]\cdots\mathrm{E}\left[e^{tX_n}\right] \\ &= \prod_{i=1}^{n} e^{\mu_i(e^t-1)} = e^{(\mu_1+\cdots+\mu_n)(e^t-1)} \quad .\end{aligned}$$

Gamma distribution

- From calculus, the integral is the Gamma function $\Gamma(\alpha)$, for $\alpha > 0$ and is a positive number.

$$\Gamma(\alpha) = \int_{y=0}^{\infty} y^{\alpha-1} e^{-y} dy \quad .$$

- If $\alpha = 1$, $\Gamma(1) = \int_{y=0}^{\infty} e^{-y} dy = 1 \quad .$
- For integral $\alpha > 1$, by integration by parts,

$$\int_0^{\infty} y^{\alpha-1} e^{-y} dy = \left[y^{\alpha-1} \int e^{-y} \right]_0^{\infty} - \int_0^{\infty} (\alpha-1) y^{\alpha-2} (-e^{-y}) dy$$

$$= 0 + (\alpha-1) \int_0^{\infty} y^{\alpha-2} e^{-y} dy = (\alpha-1)\Gamma(\alpha-1) \quad .$$

- Accordingly, for any positive integer $\alpha > 1$,

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\cdots(1)\Gamma(1) = (\alpha-1)!$$

Gamma distribution

- In the definition of the $\Gamma(\alpha)$ integral, change variable to $y = x/\beta$, where, $\beta > 0$. This gives

$$\Gamma(\alpha) = \int_{y=0}^{\infty} y^{\alpha-1} e^{-y} dy = \int_{x=0}^{\infty} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \cdot \frac{1}{\beta} dx$$

or, equivalently,

$$\frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{x=0}^{\infty} x^{\alpha-1} e^{-x/\beta} dx = 1 \quad .$$

- This is the definition of the pdf of a **Gamma distribution** with shape parameter α and scale parameter β.

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty$$

and zero when $x \leq 0$.

- The above probability distribution is sometimes denoted as $\Gamma(\alpha, \beta)$.

Exponential Distribution

- The exponential distribution is obtained from the Gamma distribution with parameters $\alpha = 1$ and $\lambda = 1/\beta > 0$ and fixed.
- pdf of exponential distribution is

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Mgf of Gamma distribution

- The mgf of a Gamma distribution with parameters α and β are as follows.

$$\mathrm{E}\left[e^{tX}\right] = \int_{x=0}^{\infty} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta+tx}$$

- Writing $e^{-x/\beta+tx} = e^{-x(1/\beta-t)}$, we change the integration variable to let $w = x(1/\beta - t)$. Assuming $1/\beta - t > 0$, or, $t < 1/\beta$, the *RHS* above is

$$\begin{aligned}\mathrm{E}\left[e^{tX}\right] &= \int_{w=0}^{\infty} \frac{1}{\Gamma(\alpha)\beta^{\alpha}(1/\beta - t)^{\alpha}} w^{\alpha-1} e^{-w} dw \\ &= \frac{1}{\Gamma(\alpha)}(1-\beta t)^{\alpha} \int_{0}^{\infty} w^{\alpha-1} e^{-w} dw \\ &= \frac{1}{(1-\beta t)^{\alpha}}\end{aligned}$$

  since the integral in the last but one step is exactly $\Gamma(\alpha)$.

- Since mgf exists, all moments $\mathrm{E}\left[X^k\right]$, for $k \geq 1$ do exist.

Moments of Gamma distribution

- Moments can be obtained from the mgf $M(t) = \mathrm{E}\left[e^{tX}\right]$: recall that $\mathrm{E}\left[X^k\right] = M^{(k)}(0)$. They can also be obtained directly from the definition.
- $M(t) = (1 - \beta t)^{-\alpha}$, $t < 1/\beta$. So,

$$\mathrm{E}[X] = M'(0) = (-\alpha)(-\beta)(1-\beta t)^{-\alpha-1}\Big|_{t=0} = \alpha\beta$$
$$\mathrm{E}\left[X^2\right] = M''(0) = \alpha\beta(\alpha+1)\beta(1-\beta t)^{-\alpha-2}\Big|_{t=0} = \alpha(\alpha+1)\beta^2$$

- Hence,

$$\mathrm{Var}[X] = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2 \quad .$$

Gamma distribution: $\mathrm{E}\left[X^k\right]$

- Gamma distribution $\Gamma(\alpha, \beta)$ satisfies an interesting property. $\mathrm{E}\left[X^k\right]$ exists for $k > -\alpha$.

$$\mathrm{E}\left[X^k\right] = \int_{x=0}^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha+k-1} e^{-x/\beta} dx$$

Multiplying and dividing by $\Gamma(\alpha+k)\beta^k$, we have,

$$\begin{aligned} &= \frac{\Gamma(\alpha+k)\beta^k}{\Gamma(\alpha)} \int_{x=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\beta^{\alpha+k}} x^{\alpha+k-1} e^{-x/\beta} dx \\ &= \frac{\Gamma(\alpha+k)\beta^k}{\Gamma(\alpha)} \ . \end{aligned}$$

Chi-square Distribution

- The chisquare distribution plays a special role. The $\chi^2(r)$ is the Gamma distribution with parameters $\alpha = r/2$ and $\beta = 2$.
- The pdf of $\chi^2(r)$ distribution is therefore,

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}, \quad x > 0$$

  and zero otherwise.
- So its mgf is

$$M(t) = (1-2t)^{-r/2}, \ \ t < \frac{1}{2} \ .$$

- Expectation is $(r/2)(2) = r$ and variance is $(r/2)(2^2) = 2r$, respectively.
- r is called the number of degrees of freedom of the chi-square distribution.

Gamma distribution: β is a scale parameter: Why?

- In the Gamma distribution $\Gamma(\alpha, \beta)$, β is called the scale parameter. Why is it?
- Let X be a r.v. having a $\Gamma(\alpha, \beta)$ distribution. Let $Y = cX$, i.e., Y is X scaled by c. The pdf of Y is

$$\begin{aligned} f_Y(y) = f_X(y/c)(1/c) &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\left(\frac{y}{c}\right)^{\alpha-1} e^{-y/(c\beta)}\frac{1}{c}dy \\ &= \frac{1}{\Gamma(\alpha)(c\beta)^\alpha} y^{\alpha-1} e^{-y/(c\beta)} \quad . \end{aligned}$$

- Hence, Y has the distribution $\Gamma(\alpha, c\beta)$.
- Some corollaries:
  - Suppose X has a Gamma distribution with $\alpha = r/2$, for some integer r and $Y = 2X/\beta$, then the distribution is $\Gamma(r/2, 2)$ which is $\chi^2(r)$ distribution.
  - β is a scale parameter, by scaling differently, this parameter changes. The shape parameter does not change by scaling.

Additive property of Gamma distribution

- Gamma distribution satisfies the (surprising!) additive property.
- ***Thm.*** Let $X_1, X_2, \ldots, X_n$ be independent random variables. Suppose that for $i = 1, 2, \ldots, n$, X_i has a $\Gamma(\alpha_i, \beta)$ distribution. Then, $X_1 + X_2 + \ldots + X_n$ has a $\Gamma(\alpha_1 + \cdots + \alpha_n, \beta)$ distribution.
- *Pf.* We will use the uniqueness of mgfs property. The mgf $M_{X_i}(t) = (1 - \beta t)^{-\alpha_i}$, $t < 1/\beta$ for $i = 1, \ldots, n$. By independence,

$$\begin{aligned} \mathrm{E}\left[e^{t(X_1 + \cdots + X_n)}\right] &= \mathrm{E}\left[e^{tX_1}\right] \mathrm{E}\left[e^{tX_2}\right] \cdots \mathrm{E}\left[e^{tX_n}\right] \\ &= \prod_{i=1}^{n} (1 - \beta t)^{-\alpha_i} = (1 - \beta t)^{-(\alpha_1 + \cdots + \alpha_n)} \quad . \end{aligned}$$

which is the mgf of $\Gamma(\alpha_1 + \cdots + \alpha_n, \beta)$ distribution.

Corollary: Additive property of chi-squared distribution

- Corollary: If $X_1, \ldots, X_n$ are independent variables, where X_i is has the $\chi^2(r_i)$ distribution, for $i = 1, 2, \ldots, n$. Let $Y = X_1 + X_2 + \cdots + X_n$. Then, Y has $\chi^2(r_1 + r_2 + \cdots + r_n)$ distribution.
- Corollary: Let X be a random variable with $\chi^2(r)$ distribution. Then,

$$\mathrm{E}\left[X^k\right] = \frac{2^k \Gamma(r/2 + k)}{\Gamma(r/2)}, \quad -r/2 < k \ .$$

This follows from the property proved for Γ distribution.

Poisson Process

- The Poisson distribution and Gamma distribution are closely related to the **Poisson process**.
- The Poisson process is used to model the number of arrivals x in an interval of time w. Examples are number of calls at a telephone switch, or packets at a network switch, or the number of alpha particles emitted by a radioactive substance that enters an observation chamber in time interval t.
- The arrival process assumes certain postulates. Let $g(x, w)$ denote the probability of x arrivals in a certain time interval w. Let $o(h)$ denotes any function g such that $\lim_{h\to 0} \frac{g(h)}{h} \to 0$.

Poisson process postulates

- The postulates are as follows. Let $h > 0$ be small. $\lambda > 0$ is a parameter.
    1. $g(1, h) = \lambda h + o(h)$, meaning that the probability of one arrival in an interval of time w is proportional to the length of the interval w, with the error term $\to 0$ as $h \to 0$.
    2. $g(2, h) + g(3, h) + \cdots + \infty = o(h)$, meaning that the probability of two or more arrivals in an interval of time h / $h \to 0$.
    3. The numbers of arrivals in non-overlapping intervals are independent.

**Poisson Process Derivations

- Recall $g(x, w)$ is the probability that there are x arrivals in a given time interval w.
- First set $x = 0$, so we are trying to obtain a closed form for $g(0, w)$ as a function of w.
- Moreover, for small interval h, the probability of one change in an interval of length h is $\lambda h + o(h) + o(h) = \lambda h + o(h)$.
- So $g(0, h) = 1 - \lambda h + o(h)$.
- By independence, $g(0, w + h) = g(0, w)g(0, h)$.
- Therefore,

$$g(0, x + h) = g(0, x)\left[1 - \lambda h - o(h)\right]$$

- Transposing and simplifying,

$$\frac{g(0, x + h) - g(0, x)}{h} = -\lambda g(0, x) - g(0, x)\frac{o(h)}{h} \quad .$$

Poisson process: derivations

- Taking the limit $h \to 0$, we have,

$$\lim_{h\to 0} \frac{g(0, w+h) - g(0, w)}{h} = -\lambda g(0, w) - g(0, w) \lim_{h\to 0} \frac{o(h)}{h} \quad .$$

- Since $\lim_{h\to 0} [o(h)/h] = 0$, therefore,

$$\frac{\partial}{\partial w} g(0, w) = -\lambda g(0, w) \quad .$$

- The solution to this differential equation is

$$g(0, w) = ce^{-\lambda w} \quad .$$

- Since, $g(0, 0) = 1$, therefore, $c = 1$ and we have the solution

$$g(0, w) = e^{-\lambda w}$$

**Poisson Process: Derivations

- We now set up an equation for $g(x, w))$, for $x > 0$. Firstly, $g(x, 0)$ is assumed to be 0, since, the probability of x arrivals in time interval of length 0 is 0.
- From Poisson postulates,

$$
\begin{aligned}
& g(x, w+h) \\
& = \mathrm{P}[\, x \text{ arrivals in time interval } (0, w+h)] \\
& = \mathrm{P}[x \text{ arrivals in interval } (0, w) \text{ and no arrivals in interval } (w, w+h)] \\
& + \mathrm{P}[x-1 \text{ arrivals in interval } (0, w) \text{ and 1 arrival in interval } (w, w+h)] \\
& = g(x, w) g(0, h) + g(x-1, w) g(1, h) \qquad \text{by independence postulate} \\
& = g(x, w)(1 - \lambda h - o(h)) + g(x-1, w)[\lambda h + o(h)]
\end{aligned}
$$

Transposing and dividing by h, we get

$$
\frac{g(x, w+h) - g(x, w)}{h} = -\lambda g(x, w) - \frac{o(h)}{h} + \lambda g(x-1, w) + g(x-1, w) \frac{o(h)}{h}
$$

Poisson Process: Derivations

- Taking the limit of $h \to 0$, we get

$$\frac{\partial}{\partial w} g(x,w) = -\lambda g(x,w) + \lambda g(x-1,w), \quad x = 1,2,3,\ldots \quad .$$

- It can be shown using mathematical induction, that the solutions to these differential equations, with boundary conditions $g(x,0) = 0$, $x = 1,2,3\ldots$ are

$$g(x,w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}$$

- For a fixed value of w, $g(x,w)$ is the Poisson pmf with parameter λw.
- That is, the number of arrivals in an interval of size w is a Poisson distribution with parameter λw.

Poisson process and Gamma distribution

- Consider the “waiting time” question. What is the waiting time for k arrivals under a Poisson process model with parameter λ.
- Let W denote the waiting time (random variable) for k arrivals.
- Its cdf is $G(w) = \mathrm{P}[W \leq w] = 1 - \mathrm{P}[W > w]$.
- The event $W > w$ is that there are fewer than k arrivals in the interval of length w, so that

$$\mathrm{P}[W > w] = \sum_{x=0}^{k-1} g(x, w) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}$$

- After some steps, including conversion of this summation into the integral, namely,

$$1 - G(w) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!} = \int_{\lambda w}^{\infty} \frac{z^{k-1} e^{-z}}{\Gamma(k)} dz$$

Poisson Process and Gamma distribution

- Now differentiating wrt w, we get

$$G'(w) = g(w) = \frac{\lambda^k w^{k-1} e^{-\lambda w}}{\Gamma(k)}, \quad 0 < w < \infty \ .$$

which is a gamma distribution with parameters $\alpha = k$ and $\beta = 1/\lambda$.

Outline

Normal Distribution

- Normal distributions provide and important family of distributions for applications and for statistical inference.
- Another motivation is the Central Limit Theorem.
- 2-Stability property is a unique and important property; widely used.

Normal distribution

- Consider the integral
$$I = \int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$
- The integral exists, since the integrand is a continuous, differentiable function which is bounded by an integrable function. (why?)
- For $z > 0$, $\frac{1}{2}(z-1)^2 \geq 0$, or $\frac{z^2}{2} > z - \frac{1}{2}$, or, $-\frac{z^2}{2} < -z + 1$.
- So for $z > 0$, $\exp\{-z^2/2\} < \exp\{-z+1\} = \exp\{-|z|+1\}$.
- For $z < 0$, $\frac{1}{2}(z+1)^2 \geq 0$, or $\frac{z^2}{2} > -z - \frac{1}{2} = |z| - \frac{1}{2}$, or, $-\frac{z^2}{2} < -|z| + 1$.
- So
$$\exp\left\{-\frac{z^2}{2}\right\} \leq \exp\{-|z|+1\}, \quad -\infty < z < \infty$$
- And,
$$\int_{-\infty}^{\infty} \exp\{-|z|+1\}\, dz = 2e \ .$$

Normal distribution

- $I = \int_{-\infty}^{\infty} \exp\left\{-z^2/2\right\} dz$.
- Note that $I > 0$ and we write I^2 as

$$I^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dxdy = \frac{1}{2\pi}\int_{0}^{\infty}\int_{0}^{2\pi} e^{-r^2/2} r\, dr\, d\theta$$

  by changing the variables (x, y) to polar coordinates (r, θ), the inverse mapping is $x = r\cos\theta$, $y = r\sin\theta$.
- The Jacobian matrix is $\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$ whose determinant is 1.
- With the change to polar coordinates, and then writing $u = r^2$, so that $du = rdr$,

$$I^2 = \frac{1}{2\pi}\int_{r=0}^{\infty}\int_{0}^{2\pi} e^{-r^2/2} rdrd\theta = \frac{1}{2\pi}\int_{u=0}^{\infty} e^{-u} du(2\pi) = 1 \quad .$$

Standard Normal Distribution

- $$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$
  is the integrand in I, is non-negative for $-\infty < z < \infty$ and integrates to 1 over $\mathbb{R}$. Hence it is a pdf.
- $f(z)$ is said to be the pdf of the **standard normal distribution.**

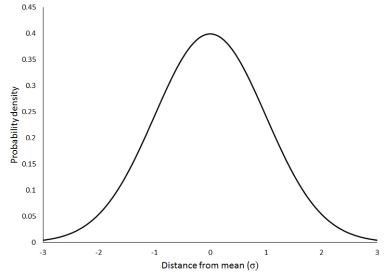

Figure: Standard normal distribution

Mgf of Standard Normal Distribution

- Moment Generating Function:
- $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. Therefore,

$$
\begin{aligned}
\mathrm{E}\left[e^{tZ}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2} + tz\right\} dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(z^2 - 2 \cdot z \cdot t + t^2\right) + \frac{t^2}{2}\right\} \\
&= \exp\{t^2/2\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(z-t)^2}{2}\right\} dz
\end{aligned}
$$

  Change variable from z to $z - t = u$, the integral is 1.
- Therefore,

$$
\mathrm{E}\left[e^{tZ}\right] = e^{t^2/2}, \qquad t \in \mathbb{R}
$$

Mean and Variance of Standard Normal Distribution

- Recall $M_Z(t) = e^{t^2/2}$, where Z has standard normal distribution.
- Then,

$$M_Z'(t) = te^{t^2/2}, \quad M_Z''(t) = e^{t^2/2} + t^2 e^{t^2/2}$$

- Therefore,

$$\mu = \mathrm{E}\left[Z\right] = M'(0) = 0$$
$$\mathrm{Var}\left[Z\right] = \mathrm{E}\left[Z^2\right] = M''(0) = 1$$

- This is typically called the $N(0, 1)$ distribution, $\mu = 0$ and $\sigma^2 = 1$.

Normal Distribution

- Define the continuous random variable as

$$X = bZ + a$$

  for $b > 0$, and Z is defined as above.
- The mapping from Z to X is 1-1 and $Z = \frac{X-a}{b}$. The Jacobian is $\left|\frac{dz}{dx}\right| = \frac{1}{b}$. Hence,

$$f_X(x) = f_Z(z(x))|J| = \frac{1}{b\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right\} \quad .$$

- From linear transformation $X = bZ + a$, $\mathrm{E}[X] = b\mathrm{E}[Z] = b \cdot 0 = 0$.
- $\mathrm{Var}[X] = b^2\mathrm{Var}[Z] = b^2$.
- Writing $\mu = \mathrm{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$, a random variable X has a normal distribution if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \quad \text{for } -\infty < x < \infty \quad .$$

Mgf of Normal Random Variable

- The random variable $X = \sigma Z + \mu$.
- Given that the mgf of Z is $e^{t^2/2}$, we have,

$$\mathrm{E}\left[e^{tX}\right] = \mathrm{E}\left[e^{t(\sigma Z+\mu)}\right] = \mathrm{E}\left[e^{t\mu} \cdot e^{(t\sigma)Z}\right]$$
$$= e^{t\mu} e^{t^2\sigma^2/2} = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$$

- The cdf of a standard normal variable Z is denoted as

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.$$

- For $X = \sigma Z + \mu$, the cdf is

$$F_X(x) = \mathrm{P}\left[X \leq x\right] = \mathrm{P}\left[\sigma Z + \mu \leq x\right] = \mathrm{P}\left[Z \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

CDF of standard normal variable

- Let Z be the standard random variable with pdf $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, $-\infty < z < \infty$.
- Clearly $f(z) = f(-z)$, for all z. f is a symmetric function. By changing the variable $y = -z$,

$$\Phi(z) = \int_{-\infty}^{z} f(z)dz = \int_{-z}^{\infty} f(-y)dy = \int_{-z}^{\infty} f(y)dy = 1 - \Phi(-z) \ .$$

or, equivalently,

$$\Phi(-z) = 1 - \Phi(z), \quad -\infty < z < \infty.$$

Normal Distribution: Remarks

- Consider the distribution $N(\mu, \sigma^2)$.
- μ is the expectation and from symmetry, is the median of the distribution. It is called the **location** parameter, where the distribution is centered.
- The standard deviation σ is called the **scale** parameter; changing its value changes the spread of the distribution.

Normal distribution and its relation to chi-squared distribution

- **Thm.** Let Z be distributed as $N(0,1)$. Then, Z^2 is distributed as $\chi^2(1)$ (which is same as Γ distribution with parameters $\alpha = \beta = 2$.)
- Pf. The cumulative probability function $F_V(v) = \mathrm{P}[V \leq v]$ is

$$
\begin{aligned}
F_V(v) &= \mathrm{P}\left[Z^2 \leq v\right] \\
&= \mathrm{P}\left[-\sqrt{v} \leq Z \leq \sqrt{v}\right] \\
&= \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} \exp\left\{-z^2/2\right\} dz \\
&= 2\int_{0}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} \exp\left\{-z^2/2\right\} dz && \text{by symmetry} \\
&= \int_{0}^{v} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{w}} e^{-w/2} dw, && \text{let } z^2 = w\ .
\end{aligned}
$$

Normal and chi-squared distributions

- This is the same as the cumulative density function for $\chi^2(1)$ distribution. The pdf is obtained by differentiating $\frac{d}{dv}F_V(v)$ which is

$$f_V(v) = \frac{d}{dv}F_V(v) = \frac{1}{\sqrt{2\pi}}v^{-1/2}e^{-v/2}, v \geq 0$$

  which is the pdf for chi-squared distribution.

- **Corollary**. If X has distribution $N(\mu, \sigma^2)$, then, $V = \left(\frac{X-\mu}{\sigma}\right)^2$ is distributed as $\chi^2(1)$.

Stability of Normal Distributions

- **Thm.** Suppose $X_1, X_2, \ldots, X_n$ are independent random variables such that for $i = 1, 2, \ldots, n$, X_i has $N(\mu_i, \sigma_i^2)$ distribution. Let $Y = a_1 X_1 + \cdots + a_n X_n$, where, the a_i's are constants. Then Y has the distribution $N(a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)$.
- $$\begin{aligned} M_Y(t) &= \mathrm{E}\left[\exp\left\{t(a_1X_1 + \ldots + a_nX_n)\right\}\right] \\ &= \mathrm{E}\left[\exp\{ta_1X_1\}\right] \cdot \mathrm{E}\left[\exp\{ta_2X_2\}\right] \cdot \cdots \cdot \mathrm{E}\left[\exp\{ta_nX_n\}\right], \text{by independ.} \\ &= \prod_{i=1}^{n} \exp\left\{ta_i\mu_i + (1/2)t^2a_i^2\sigma_i^2\right\} \\ &= \exp\left\{t(a_1\mu_1 + \ldots + a_n\mu_n) + \frac{t^2}{2}(a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2)\right\} \end{aligned}$$

  which is the mgf of a $N(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$ distribution.
- By uniqueness of mgf, Y is distributed as $N(a_1\mu_1 + \ldots + a_n\mu_n, a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2)$.

Stability property: Corollary

- **Corollary**. Let $X_1, X_2, \ldots, X_n$ be iid random variables with common distribution $N(\mu, \sigma^2)$. Let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.
  Then, $\bar{X}$ has $N(\mu, \sigma^2/n)$ distribution.

Multivariate normal distribution

- Let Z be the random vector $\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$
- Each of the Z_i's have $N(0,1)$ normal distribution.
- The pdf of $\mathbf{Z}$, from first principles by independence is,

$$\begin{aligned} f_{\mathbf{Z}}(z_1, z_2, \ldots, z_n) &= f_{Z_1}(z_1) f_{Z_2}(z_2) \cdots f_{Z_n}(z_n) \\ &= \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-z_2^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-z_n^2/2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(z_1^2 + z_2^2 + \cdots + z_n^2)} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\mathbf{z}^T \mathbf{z}/2} \end{aligned}$$

Multivariate normal distribution: Mean and Covariance

- We have used

$$\mathbf{z}^T\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = z_1^2 + \cdots + z_n^2 \ .$$

- Each of the Z_i's have 0 mean, therefore,

$$\mathrm{E}\left[\mathbf{Z}\right] = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}.$$

- $\mathrm{Var}\left[Z_i\right] = 1$, for $i = 1, \ldots, n$, and for $i \neq j$, by independence, $\mathrm{Cov}\left(Z_i, Z_j\right) = 0$. Therefore,

$$\mathrm{Cov}\left(\mathbf{Z}\right) = \mathbf{I}_n.$$

Mgf of multivariate normal distribution

- The mgf of each Z_i as a function of t_i is $M(t_i) = e^{t_i^2/2}$.
- The mgf of $\mathbf{Z}$ is, by independence of the Z_i's,

$$
\begin{aligned}
\mathrm{E}\left[e^{t_1 Z_1 + \cdots + t_n Z_n}\right] &= \mathrm{E}\left[\prod_{i=1}^{n} e^{t_i Z_i}\right] \\
&= \prod_{i=1}^{n} \mathrm{E}\left[e^{t_i Z_i}\right] \\
&= \prod_{i=1}^{n} e^{t_i^2/2} \\
&= e^{(t_1^2 + t_2^2 + \cdots + t_n^2)/2} \qquad \text{for all } t_1, \ldots, t_n \in \mathbb{R}.
\end{aligned}
$$

- In vector notation, we may write $\mathbf{t}$ to be the n-dimensional vector $\mathbf{t}^T = (t_1, t_2, \ldots, t_n)^T$. Above is abbreviated as

$$
\mathrm{E}\left[e^{\mathbf{t}^T \mathbf{z}}\right] = e^{\mathbf{t}^T \mathbf{t}/2} \quad .
$$

Multivariate normal distribution

- We say that $\mathbf{Z}$ has a mutlivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_n$.
- The multivariate distribution is denoted as $N_n(\mathbf{0}, \mathbf{I})$.

Positive Definiteness: review

- By spectral theorem of linear algebra, every symmetric matrix A has a full set of eigenvectors, which are orthonormal to each other.
- Let $U = \begin{bmatrix} U_1 & U_2 & \cdots & U_n \end{bmatrix}$ denote the eigenvector matrix of A whose columns U_i's are the orthogonal eigenvectors.
- The ith eigenvector U_i corresponds to the eigen value λ_i, that is,

$$AU_i = \lambda_i U_i$$

- In matrix form,

$$\begin{aligned} AU = A\begin{bmatrix} U_1 & U_2 & \cdots & U_n \end{bmatrix} &= \begin{bmatrix} \lambda_1 U_1 & \lambda_2 U_2 & \cdots & \lambda_n U_n \end{bmatrix} \\ &= \begin{bmatrix} U_1 & U_2 & \cdots & U_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} = U\Lambda \end{aligned}$$

Covariance matrix Σ

- Λ is the diagonal matrix $\Lambda_{ii} = \lambda_i$.

$$AU = U\Lambda \quad \text{or, } AU\Lambda U^T$$

  since U is an orthogonal matrix; so $U^{-1} = U^T$.

- The covariance matrix Σ is positive semi-definite– all eigen-values $\lambda_1, \ldots, \lambda_n$ are non-negative.

$$\Sigma = U\Lambda U^T \quad .$$

- Since eigenvalues are non-negative , define the diagonal matrix

$$\Lambda^{1/2} = \begin{bmatrix} \lambda^{1/2} & & \\ & \ddots & \\ & & \lambda_n^{1/2} \end{bmatrix}$$

Covariance Matrix

- A "square root" of the matrix Σ is defined as

$$\Sigma^{1/2} = U\Lambda^{1/2}U^T$$

- To see that it is a "square root",

$$\Sigma^{1/2}\Sigma^{1/2} = (U\Lambda^{1/2}U^T)(U\Lambda^{1/2}U^T) = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = \Sigma$$

- $\Sigma^{1/2}$ is also symmetric and positive semi-definite.
- Assuming Σ is Positive-definite (all λ_i's are positive),

$$\Sigma^{-1} = (U\Lambda U^T)^{-1} = (U^T)^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T \quad .$$

- Likewise similarly,

$$(\Sigma^{1/2})^{-1} = U\Lambda^{-1/2}U^T \quad .$$

- $(\Sigma^{1/2})^{-1}$ is denoted as $\Sigma^{-1/2}$.

Multivariate normal distribution: general form

- Let Σ be an n by n positive semi-definite matrix. Let $\mu = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \end{bmatrix}^T$ be an n-dimensional vector.
- Let Z be a random vector with $N(\mathbf{0}, I_n)$ distribution.
- Define
$$X = \Sigma^{1/2} Z + \mu \quad .$$
- By linearity of expectation,
$$\mathrm{E}\left[X\right] = \mathrm{E}\left[\Sigma^{1/2} Z + \mu\right] = \Sigma^{1/2} \mathrm{E}\left[Z\right] + \mu = \Sigma^{1/2} \cdot \mathbf{0} + \mu = \mu \quad .$$
-
$$\begin{aligned} \mathrm{Cov}\left(X\right) &= \mathrm{Cov}\left(\Sigma^{1/2} Z + \mu\right) = \mathrm{Cov}\left(\Sigma^{1/2} Z\right) \\ &= \Sigma^{1/2} \mathrm{Cov}\left(Z\right) \left(\Sigma^{1/2}\right)^T = \Sigma^{1/2} I_n \Sigma^{1/2} = \Sigma \quad . \end{aligned}$$

  since $\Sigma^{1/2}$ is a symmetric matrix.

General form

- The mgf of X is calculated as follows. Let t be the n-dimensional vector $t = (t_1, \ldots, t_n)$.

$$
\begin{aligned}
M_X(t) &= \mathrm{E}\left[\exp\{t^T X\}\right] \\
&= \mathrm{E}\left[\exp\{t^T(\Sigma^{1/2} Z + \mu)\}\right] \\
&= \exp\{t^T\mu\}\mathrm{E}\left[\exp\{t^T\Sigma^{1/2} Z\}\right] \\
&= \exp\{t^T\mu\}\mathrm{E}\left[\exp\{(\Sigma^{1/2} t)^T Z\}\right] \\
&= \exp\{t^T\mu\}\exp\{(\Sigma^{1/2} t)^T(\Sigma^{1/2} t)/2\} \qquad \text{by mgf of } Z \\
&= \exp\{t^T\mu + (1/2)t^T\Sigma t\}
\end{aligned}
$$

Mgf

- The key step in the previous calculation is that of $\mathrm{E}\left[\exp\{(\Sigma^{1/2}t)^T Z\}\right]$.
- It was earlier proved that for any $t \in \mathbb{R}^n$, $\mathrm{E}\left[\exp\{t^T Z\}\right] = \mathrm{E}\left[t^T t/2\right]$.
- Let $s = (\Sigma^{1/2}t)^T$. Hence,

$$\mathrm{E}\left[\exp\{s^T Z\}\right] = \exp\{s^T s/2\} = \exp\{(\Sigma^{1/2}t)^T(\Sigma^{1/2}t)/2\}.$$

- Note the validity of $M_Z(t) = e^{t^T t/2}$; this holds for all $t \in \mathbb{R}^n$ and allows the above inference.

Mgf of General multivariate normal distribution

- **Definition.** An n-dimensional random vector X is said to have a mutlivariate normal distribution if its mgf is

$$M_X(t) = \exp\{t^T\mu + (1/2)t^T\Sigma t\}, \quad \text{for all } t \in \mathbb{R}^n,$$

where, Σ is a symmetric positive semi-definite matrix and $\mu \in \mathbb{R}^n$. The distribution is denoted as $N(\mu, \Sigma)$.

Pdf of multivariate normal distribution

- Let Σ be a positive definite matrix. Hence it is invertible and so is $\Sigma^{1/2}$.
- For $X = \Sigma^{1/2}Z + \mu$, the inverse mapping is well-defined,

$$Z = \Sigma^{-1/2}(X - \mu).$$

- Let W be a random vector of n variables and let $V = AW$, where, A is an n by n matrix of constants. Then,

$$\frac{\partial V_i}{\partial W_j} = \frac{1}{\partial W_j}\sum_{k=1}^{n} A_{ik}W_k = A_{ij}, \quad 1 \leq i \leq n, 1 \leq j \leq n$$

- Hence, the Jacobian matrix of V w.r.t. W is A.
- Applying this to the inverse mapping $Z = \Sigma^{-1/2}(X - \mu)$, the Jacobian matrix is $\Sigma^{-1/2}$ and hence,

$$|\det J| = \det(\Sigma^{-1/2}) = \frac{1}{|\det \Sigma|^{1/2}} .$$

by property of determinants.

Pdf of multivariate normal distributions

- $X = \Sigma^{1/2} Z + \mu$, and the inverse mapping is $Z = \Sigma^{-1/2}(X - \mu)$.
- Hence, by transforming Z to X, we have,

$$
\begin{aligned}
f_X(x) &= f_Z(z(x)) |\det J| \\
&= \frac{1}{(2\pi)^{n/2} |\det \Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} ((\Sigma^{-1/2}(x-\mu)))^T \Sigma^{-1/2}(x-\mu) \right\} \\
&= \frac{1}{(2\pi)^{n/2} |\det \Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T (\Sigma^{-1/2})^T (\Sigma^{1/2}(x-\mu)) \right\} \\
&= \frac{1}{(2\pi)^{n/2} |\det \Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}
\end{aligned}
$$

- Note that the Σ is symmetric, positive definite, and so $(\Sigma^{-1/2})^T \Sigma^{-1/2} = \Sigma^{-1/2} \Sigma^{-1/2} = \Sigma^{-1}$.

Linear Transformation of a Multivariate Normal Variable

- **Property**. Let X have a $N_n(\mu, \Sigma)$ distribution. Let $Y = AX + b$, where, A is an m by n matrix and $b \in \mathbb{R}^m$. Then Y has distribution $N_m(A\mu + b, A\Sigma A^T)$.
- Proof is via calculating the mgf of Y, $M_Y(t)$. (Here, $t = (t_1, \dots, t_m)^T$). We use that $M_X(s) = \exp\left\{s^T\mu + (1/2)s^T\Sigma s\right\}$, for all $s \in \mathbb{R}^n$.

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}\left[e^{t^T Y}\right] \\
&= \mathrm{E}\left[\exp\{t^T(AX + b)\}\right] \\
&= \exp\{t^T b\}\mathrm{E}\left[\exp\{t^T AX\}\right] \\
&= \exp\{t^T b\}\mathrm{E}\left[\exp\{(A^T t)^T X\}\right] \\
&= \exp\{t^T b\}\exp\{(A^T t)^T\mu + (1/2)(A^T t)^T\Sigma A^T t\}, \quad s = A^T t \\
&= \exp\{t^T(b + A\mu) + (1/2)t^T A\Sigma A^T t\}
\end{aligned}
$$

which is the mgf of an $N_m(A\mu + b, A\Sigma A^T)$ distribution.

Notes

- (*Re-statement*: Let X have a $N_n(\mu, \Sigma)$ distribution. Let $Y = AX + b$, where, A is an m by n matrix and $b \in \mathbb{R}^m$. Then Y has a $N_m(A\mu + b, A\Sigma A^T)$.
- Note that if A has rank m, then, its pdf can be found as before. (Exercise!).
- If A has rank $r < m$, then find its pdf - Exercise! Note that the pdf is defined for only some specific set of r variables of Y, the remaining $m - r$ variables are linear functions of those r variables.
- This latter case arises if $m > n$.

Another application: Marginal Distribution

- Let X_1 be any sub-vector of X of dimension $m < n$.
- Rerrange the variables in X (and accordingly rearrange the mean and covariance matrix), and write

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where X_2 are the remaining $n - m$ variables of X and of dimension $n - m$.
- Accordingly partition the mean and covariance matrix of X:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where, $\Sigma_{11} = \mathrm{Cov}\,(X_1)$ and is $m \times m$, $\Sigma_{12} = \mathrm{Cov}\,(X_1, X_2)$ is $m \times (n - m)$, etc..

Marginal Distribution

- Let

$$A = \begin{bmatrix} I_m & 0_{m\times(n-m)} \end{bmatrix}$$

- Then, $X_1 = AX$.
- Applying the earlier theorem, we get the corollary.
- *Corollary.* Suppose X has the $N_n(\mu, \Sigma)$ distribution, partitioned as given earlier. Then, X_1 has a $N_m(\mu_1, \Sigma_{11})$ distribution.
- I.e., *any marginal distribution of X is also normal, and its mean and variance are those associated with that partial vector* (only).

Corollary: Rotational Invariance of Normal Distributions

- Let Z have an $N_n(0, I)$ distribution.
- Let A be an m by n matrix whose rows are orthogonal ($AA^T = I_m$).
- Then, A has the distribution $N_m(A \cdot 0, AIA^T)$. Since, $AIA^T = AA^T = I_m$.
- Hence AZ has distribution $N_m(0, I_m)$.
- As a special case, if A is n by n orthogonal matrix, then, AZ has distribution $N_n(0, I_n)$ and is identically distributed as Z.
- Denoting $Y = AZ$, $\Sigma = \mathrm{Cov}\,(Y) = \mathrm{Cov}\,(AX) = A\mathrm{Cov}\,(Z)\,A^T = A \cdot I \cdot A^T = I$.
- The notation in the exponent of the pdf of Y would be

$$\exp\{-(1/2)(A^{-1}z)^T(A^{-1}z)\} = \exp\left\{-(1/2)\left\|A^{-1}z\right\|_2^2\right\}$$

Rotational invariance

- The matrix expression is

$$\left\|A^{-1}z\right\|_2^2 = z^T(A^{-1})^T A^{-1} z = z^T A A^T z = z^T I z = \|z\|_2^2 \quad .$$

- I.e., Distribution of AZ is identical to that of Z under any orthogonal transformation A: “rotational invariance”.

Normal Distribution: Uncorrelated $\Leftrightarrow$ Independent.

- The following is an important property of normal distributions.
- Let X have a $N_n(\mu, \Sigma)$ distribution and let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, where, X_1 and X_2 partition X into m variables and $n - m$ variables.
- *Suppose* $\mathrm{Cov}\,(X_1, X_2) = \Sigma_{12} = \Sigma_{21}^T = 0_{m\times(n-m)}$.
- *Then, X_1 and X_2 are independent.*
- Converse is obviously true.

Uncorrelated implies Independence

- Corresponding to the partition of X into X_1 and X_2, partition t into sub-vectors t_1 and t_2.
- We calculate $M_X(t) = M_{X_1,X_2}(t_1, t_2)$.

$$M_{X_1,X_2}(t_1, t_2) = \exp\left\{t_1^T \mu_1 + t_2^T \mu_2\right\} \cdot \exp\left\{t^T \Sigma t\right\}$$

- By uncorrelatedness, $\Sigma_{12} = 0$ and $\Sigma_{21} = 0$. Hence

$$\begin{aligned} t^T \Sigma t &= \begin{bmatrix} t_1^T & t_2^T \end{bmatrix} \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \\ &= t_1^T \Sigma_{11} t_1 + t_2^T \Sigma_{22} t_2 \end{aligned}$$

  Therefore,

$$\begin{aligned} M_{X_1,X_2}(t_1, t_2) &= \exp\left\{t_1^T \mu_1 + t_1^T \Sigma_{11} t_1\right\} \exp\left\{t_2^T \mu_2 + t_2^T \Sigma_{22} t_2\right\} \\ &= M_{X_1}(t_1) M_{X_2}(t_2) \end{aligned}$$

  and hence X_1, X_2 are independent.

*Conditional Distribution of $X_1 \mid X_2$

- Let X be an n-dimensional normal variate $N(\mu, \Sigma)$ and is partitioned as $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Assume X_1 is m-dimensional.
- Question: what is the distribution of $X_1 \mid X_2$? Proof is in two steps.
- Step 1: Recall $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Consider the transformation:

$$\begin{bmatrix} W \\ X_2 \end{bmatrix} = \begin{bmatrix} I_m & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-m} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

*Conditional Distribution

- Hence,

$$\mathrm{Cov}\left(\begin{bmatrix} W \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} I_m & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-m} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_m & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{n-m} \end{bmatrix}$$
$$= \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & \Sigma_{22} \end{bmatrix}$$

- From earlier discussion and theorems, the random vectors W and X_2 are therefore independent.
- Hence, $W \mid X_2 = x_2$ has the same distribution as the marginal distribution of W, which is
$$N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$
- Given $X_2 = x_2$, $W + \Sigma_{12}\Sigma_{22}^{-1}X_2$ has the distribution
$$N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

*Conditional Distribution

- This holds for all $x_2 \in \mathbb{R}^{n-m}$, so $X_1 = W + \Sigma_{12}\Sigma_{22}^{-1}X_2$ conditioned on X_2 has the distribution

$$N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}X_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

- (An interesting and surprising result!). $X_1 \mid X_2$ has a normal distribution with the above mentioned parameters.

Remarks

- We have proved earlier that if Z is $N(0,1)$, then, Z^2 is $\chi^2(1)$.
- We also know that the sum of n iid $\chi^2(1)$ variables is $\chi^2(n)$.
- If Z is n-dimensional normal variate $N_n(0, I)$, then, what is the distribution of $Z^T Z = \|Z\|^2$?
- Since $Z_1, \ldots, Z_n$ are independent,

$$Z^T Z = Z_1^2 + \ldots + Z_n^2 .$$

- Also, $Z_i^2 \sim \chi^2(1)$, $i = 1, 2, \ldots, n$.
- Hence, $Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$.

Remarks

- Generalizing, let X be distributed as $N_n(\mu, \Sigma)$. Let Σ be positive definite.
- We can define
$$Z = \Sigma^{-1/2}(X - \mu) \ .$$
- Then, Z has distribution $N_n(0, I)$, since,
  1. $\mathrm{E}[Z] = \Sigma^{-1/2}\mathrm{E}[X - \mu] = \Sigma^{1/2} \cdot 0 = 0$.
  2. $\mathrm{Cov}(Z) = \Sigma^{-1/2}\Sigma(\Sigma^{-1/2})^T = I$.
- By the argument earlier, $Z^T Z = \|Z\|_2^2$ has $\chi^2(n)$ distribution.
- Hence, $(X - \mu)^T\Sigma^{-1}(X - \mu) = Z^T Z = \left\|\Sigma^{-1/2}(X - \mu)\right\|_2^2$ has $\chi^2(n)$ distribution!

Total Variation

- Let the random vector X have distribution $N_n(\mu, \Sigma)$.
- **Definition.** The total variation (TV) of X is defined as the sum of the variances of its components. That is,

$$TV(X) = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = \mathrm{Tr}\ \Sigma \ \ .$$

- Write the eigen decomposition of Σ as

$$\Sigma = U \Lambda U^T$$

- For purposes of this discussion, we assume

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n > 0$$

  and the vectors in U are rearranged accordingly.

Principal components

- Define the linear mapping
$$Y = U^T(X - \mu) \ .$$
- $\mathrm{E}[Y] = U^T \mathrm{E}[X - \mu] = U^T \cdot 0 = 0.$
-
$$\mathrm{Cov}(Y) = \mathrm{Cov}\left(U^T X\right) = U^T \mathrm{Cov}(X)\, U = U^T (U \Lambda U^T) U = \Lambda$$
- So, Y is distributed as $N_n(0, \Lambda)$.
- The components random vectors of Y are all mutually independent.
- The random vector Y is called the **vector of principal components.**

Total Variation of Y

- The total variation of Y is

$$\mathrm{Tr}\ \Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

- How is it related to the total variation of X?

$$TV(X) = \mathrm{Tr}\ \Sigma = \mathrm{Tr}\ U\Lambda U^T = \mathrm{Tr}\ \Lambda U^T U = \mathrm{Tr}\ \Lambda = TV(Y)$$

- We therefore have the following property.

$$TV(X) = \sum_{i=1}^{n} \sigma_i^2 = \sum_{i=1}^{n} \lambda_i = TV(Y)$$

- In general, if V is any orthogonal transformation, then, the total variation of VX is the same as that of X:

$$TV(VX) = \mathrm{Tr}\ V\Sigma V^T = \mathrm{Tr}\ \Sigma V^T V = \mathrm{Tr}\ \Sigma = TV(X)\ .$$

Linear combination of X with maximum variance

- Consider the following question. Given X with distribution $N_n(\mu, \Sigma)$, find a unit vector v such that $v^T X$ has maximum variance.
- Since $\Sigma = U \Lambda U^T$, we can write any vector v as $v = Uw$, uniquely. Then, $\|v\| = \|w\| = 1$.

$$\mathrm{Var}\left[v^T X\right] = v^T \Sigma v = (Uw)^T U \Lambda U^T (Uw) = w^T \Lambda w = \sum_{i=1}^{n} \lambda_i w_i^2 \ .$$

- Since, $\|w\|_2^2 = 1 = \sum_{i=1}^n w_i^2$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$,
- $\mathrm{Var}\left[v^T X\right]$ is maximum when $w_1 = 1$ and $w_2 = \cdots = w_n = 0$ and equals λ_i.
- Thus, $w = e_1 = \begin{bmatrix} 1 & 0 & \vdots & 0 \end{bmatrix}$, and $v = Ue_1 = U_1$, the first eigen vector.
- Thus,

$$\mathrm{Var}\left[v^T X\right] \leq \mathrm{Var}\left[U_1^T X\right] = \lambda_1 \ .$$

Principal components

- Analogously, U_2 is the second principal component, since $U_2^T X$ has the largest variance among all $v^T X$, where, v is a unit vector and $v \perp U_1$. (Proof is similar).
- Similarly, $U_3, \ldots U_n$ are the third, fourth and successive principal component.