STA347 Probability I

Gun Ho Jang

Update on December 3, 2018

Note: This note is prepared for STA347. There might be numerous fault arguments/statements/typos. If you spot one, please contact the instructor or you may look up references which may contain errors too.

Applications

Normal Distributions

A multivariate normal distribution $Z = (Z_1, \ldots, Z_k) \sim N_k(\mu, \Sigma)$ has the probability density function

$$\mathrm{pdf}_Z(z) = \mathrm{pdf}_{Z_1,\ldots,Z_k}(z_1, \ldots, z_k) = |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu))$$

and the moment generating function

$$\begin{aligned}\mathrm{mgf}_Z(t) = \mathrm{mgf}_{Z_1,\ldots,Z_k}(t_1, \ldots, t_k) = \mathbb{E}[e^{t^\top Z}] &= \int e^{t^\top z} \cdot |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu))\, dz \\ &= e^{t^\top \mu} \int e^{t^\top y} \cdot |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2} y^\top \Sigma^{-1} y)\, dy = e^{t^\top \mu + t^\top \Sigma t/2} \int |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2}(y-\Sigma t)^\top \Sigma^{-1}(y-\Sigma t))\, dy \\ &= e^{t^\top \mu + t^\top \Sigma t/2}.\end{aligned}$$

Lemma 94. Assume $Z \sim N_k(\mu, \Sigma)$

(a) For any $m \times k$ matrix A and $m \times 1$ vector b, $AZ + b \sim N_m(A\mu + b, A\Sigma A^\top)$.

(b) $\mathbb{E}(Z) = \mu$ and $\mathbb{V}\mathrm{ar}(Z) = \mathbb{E}[(Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^\top] = \Sigma$.

(c) Two random variables Z_i and Z_j are independent if and only if they are uncorrelated $\mathrm{Cov}(Z_i, Z_j) = 0$.

Proof. (a) Let $Y = (Y_1, \ldots, Y_m) = AZ + b$. The moment generating function of Y is

$$\begin{aligned}\mathrm{mgf}_Y(s) &= \mathbb{E}[e^{s^\top Y}] = \mathbb{E}[e^{s^\top b + s^\top AZ}] = e^{s^\top b}\mathbb{E}[e^{(A^\top s)^\top Z}] = e^{s^\top b}\mathrm{mgf}_Z(A^\top s) = e^{s^\top b}\exp((A^\top s)^\top \mu + (A^\top s)^\top \Sigma(A^\top s)/2) \\ &= \exp(s^\top(A\mu + b) + s^\top(A\Sigma A^\top)/2) \sim N_m(A\mu + b, A\Sigma A^\top).\end{aligned}$$

(b) From the moment equation,

$$\mathbb{E}[Z] = \frac{\partial}{\partial t}\mathrm{mgf}_Z(t)|_{t=0} = \frac{\partial}{\partial t}\Big(\exp(t^\top\mu + \frac{1}{2}t^\top\Sigma t)\Big)\Big|_{t=0} = (\mu + \Sigma t)\Big(\exp(t^\top\mu + \frac{1}{2}t^\top\Sigma t)\Big)\Big|_{t=0} = \mu$$

and

$$\begin{aligned}\mathbb{E}[ZZ^\top] &= \frac{\partial^2}{\partial t\partial t^\top}\mathrm{mgf}_Z(t)|_{t=0} = \frac{\partial}{\partial t^\top}(\mu + \Sigma t)\Big(\exp(t^\top\mu + \frac{1}{2}t^\top\Sigma t)\Big)\Big|_{t=0} = [(\mu + \Sigma t)(\mu + \Sigma t)^\top + \Sigma]\Big(\exp(t^\top\mu + \frac{1}{2}t^\top\Sigma t)\Big)\Big|_{t=0} \\ &= \mu\mu^\top + \Sigma.\end{aligned}$$

Therefore

$$\mathbb{V}\mathrm{ar}(Z) = \mathbb{E}\Big[(Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^\top\Big] = \mathbb{E}\Big[ZZ^\top\Big] - \mathbb{E}(Z)\mathbb{E}(Z)^\top = \mu\mu^\top + \Sigma - \mu\mu^\top = \Sigma.$$

(c) Considering $Z_i = e_i^\top Z$, the marginal distribution of Z_i is $Z_i = e_i^\top Z \sim N(\mu_i, \Sigma_{ii})$. Similarly $Z_j \sim N(\mu_j, \Sigma_{jj})$. Let $A = (e_i, e_j)$ so that $A_{i,1} = A_{j,2} = 1$ and $A_{i',j'} = 0$ otherwise. Then $(Z_i, Z_j) \sim N_2((\mu_i, \mu_j), \begin{pmatrix}\Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj}\end{pmatrix})$ and the joint moment generating function is

$$\begin{aligned}\mathrm{mgf}_{Z_i,Z_j}(s,t) &= \exp((s,t)^\top(\mu_i, \mu_j) + (s,t)^\top\Sigma_{(i,j),(i,j)}(s,t)/2) = \exp(s\mu_i + t\mu_j + \Sigma_{ii}s^2/2 + \Sigma_{jj}t^2/2 + \Sigma_{ij}st) \\ &= \mathrm{mgf}_{Z_i}(s)\mathrm{mgf}_{Z_j}(t)\cdot\exp(\Sigma_{ij}st).\end{aligned}$$

Hence Z_i and Z_j are independent if and only if $\Sigma_{ij} = 0$ or $\mathrm{Cov}(Z_i, Z_j) = 0$. □

Theorem 95. Suppose $X_1, X_2, \ldots$ is a random sample from $N(\mu, \sigma^2)$. Then $\bar{X}_n = (X_1 + \cdots + X_n)/n$ and $S_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are independent and converge to the true parameters μ and σ^2.

Proof. Assume $\mu = 0$ and $\sigma^2 = 1$. Let $V_n = (n-1)S_n = \mathbf{X}_n^\top(I_n - \mathbf{1}_n\mathbf{1}_n^\top/n)\mathbf{X}_n$ for simplicity where

$\mathbf{X}_n = (X_1, \ldots, X_n)^\top$. The joint characteristic function of $\bar{X}_n, V_n$ is

$$
\begin{aligned}
\mathrm{mgf}_{\bar{X}_n, V_n}(s,t) &= \mathbb{E}[\exp(s\bar{X}_n + tV_n)] = \mathbb{E}[\exp((s/n)\mathbf{1}_n^\top \mathbf{X}_n + \mathbf{X}_n^\top (tI_n - (t/n)\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}_n)] \\
&= (2\pi)^{-n/2} \int \exp((s/n)\mathbf{1}_n^\top \mathbf{x} + \mathbf{x}^\top (tI_n - (t/n)\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{x} - \mathbf{x}^\top \mathbf{x}/2)\, d\mathbf{x} \\
&= \left|(1-2t)I_n + 2t/n\mathbf{1}_n\mathbf{1}_n^\top\right|^{1/2} \exp[(s/n)^2 \mathbf{1}_n^\top [(1-2t)I_n + 2t/n\mathbf{1}_n\mathbf{1}_n^\top)^{-1}\mathbf{1}_n/2] \\
&= (1-2t)^{-(n-1)/2} \exp(s^2/(2n)).
\end{aligned}
$$

Hence $\bar{X}_n \sim N(\mu, \sigma^2/n)$ and $V_n/\sigma^2 \sim \mathrm{gamma}((n-1)/2, 1/2)$ are independent.

The strong law of large numbers imply $\bar{X}_n \longrightarrow \mu$ almost surely and

$$
S_n = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \longrightarrow \mathbb{E}[X_1^2] - \mu^2 = \sigma^2 \quad \text{almost surely.}
$$

In the above, the strong law of large numbers and continuous mapping theorem are used.

Remind $\bar{X}_n = \mathbf{1}_n^\top \mathbf{X}$. Considering the second and fourth moments are finite, $\mathbb{E}[|\bar{X}_n - \mu|^2] \to 0$ and $\mathbb{E}[(S_n - \sigma^2)] \to 0$. □

It is easy to see that $\bar{X}_n \longrightarrow \mu$ in probability and $\sqrt{n}(\bar{X}_n - \mu) \sim N(0, \sigma^2)$.

Exercise 46. Let $X_1, X_2, \ldots$ be a random sample from a distribution having finite variance. Show that $\bar{X}_n$ and S_n are unbiased estimator of μ and σ^2.

Linear Regression

Consider a linear regression model

$$
Y_i = \alpha + \beta x_i + \varepsilon_i
$$

where $\varepsilon_i \sim i.i.d.$ with $\mathbb{E}(\varepsilon_n) = 0$, $\sigma^2 = \mathbb{V}\mathrm{ar}(\varepsilon_n)$ and $\mathbb{E}(\varepsilon_n^4) < \infty$ and covariates $x_1, x_2, \ldots$ are treated as known.

When $Y_1, \ldots, Y_n$ are observed, the parameters α, β can be estimated using the least square method given by

$$
(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{j=1}^n [Y_j - (\alpha + \beta x_j)]^2.
$$

The least square estimators are easily solvable using differentials or perfect squares, that is,

$$\widehat{\beta} = S_{XY}/S_{XX}$$
$$\widehat{\alpha} = \bar{Y}_n - \widehat{\beta}\bar{x}_n$$

where $S_{XY} = \sum_{j=1}^n (x_j - \bar{x}_n)(Y_j - \bar{Y}_n)$ and $S_{XX} = \sum_{j=1}^n (x_j - \bar{x}_n)^2$. By plugging-in $Y_j = \alpha + \beta x_j + \varepsilon_j$,

$$\widehat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{j=1}^n (x_j - \bar{x}_n)(\beta(x_j - \bar{x}_n) + \varepsilon_j - \bar{\varepsilon}_n)}{\sum_{j=1}^n (x_j - \bar{x}_n)^2} = \beta + \frac{\sum_{j=1}^n (x_j - \bar{x}_n)\varepsilon_j}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

The convergence of $\widehat{\beta}$ is equivalent to the convergence of $\widehat{\alpha}$, that is,

$$\widehat{\alpha} = \bar{Y}_n - \widehat{\beta}\bar{x}_n = \alpha + \beta\bar{x}_n + \bar{\varepsilon}_n - \widehat{\beta}\bar{x}_n = \alpha - (\widehat{\beta} - \beta)\bar{x}_n + \bar{\varepsilon}_n$$

converges to α in the same mode to $\widehat{\beta} \to \beta$ because $\bar{\varepsilon}_n \to 0$ both in L^2 and almost surely.

By taking expectation, $\widehat{\beta}$ is an unbiased estimator of β. The variance is

$$\mathbb{V}\text{ar}(\widehat{\beta}) = \sum_{j=1}^n \Big(\frac{x_j - \bar{x}_n}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\Big)^2 \mathbb{V}\text{ar}(\varepsilon_j) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Using Chebychev's inequality, $\widehat{\beta}$ converges to β in probability if $\sum_{j=1}^n (x_j - \bar{x}_n)^2$ diverges. Due to the unbiasedness, $\widehat{\beta} \longrightarrow \beta$ in L^2 sense, too.

Besides, if $n^{-1}\sum_{j=1}^n (x_j - \bar{x}_n)^2 \to \tau^2 > 0$, then $\widehat{\beta} \xrightarrow{a.s.} \beta$. For a proof, note that $\sum_{j=1}^n (x_j - \bar{x}_n)^2 > (1-\epsilon)n\tau^2$ for $0 < \epsilon < 1$ and sufficiently large n. Then

$$\begin{aligned}
\mathbb{E}[(\widehat{\beta} - \beta)^4] &= \mathbb{E}\Big[\frac{1}{S_{XX}}\sum_{j=1}^n (x_j - \bar{x}_n)\varepsilon_j\Big]^4 = \frac{1}{S_{XX}^4}\sum_{j \neq k=1}^n (x_j - \bar{x}_n)^2 (x_k - \bar{x}_n)^2 \mathbb{E}[\varepsilon_j^2\varepsilon_k^2] + \frac{1}{S_{XX}^4}\sum_{j=1}^n (x_j - \bar{x}_n)^4 \mathbb{E}[\varepsilon_j^4] \\
&= \frac{1}{S_{XX}^4} S_{XX}^2 [\sigma_\varepsilon^2]^2 + \frac{1}{S_{XX}^4}\sum_{j=1}^n (x_j - \bar{x}_n)^4 [\mathbb{E}(\varepsilon_j^4) - \sigma_\varepsilon^4] \le \frac{\sigma_\varepsilon^4}{S_{XX}^2} + \frac{1}{S_{XX}^2}[\mathbb{E}(\varepsilon_j^4) - \sigma_\varepsilon^4] \\
&\le \frac{\mathbb{E}[\varepsilon^4]}{(1-\epsilon)^2\tau^4 n^2}.
\end{aligned}$$

and by applying Theorem 67 based on

$$\sum_{n=1}^\infty P(|\widehat{\beta} - \beta| > \epsilon) \le \sum_{n=1}^\infty \frac{1}{\epsilon^4}\mathbb{E}[|\widehat{\beta} - \beta|^4] \le \frac{\mathbb{E}[\varepsilon^4]}{\epsilon^4(1-\epsilon)^2\tau^4}\sum_{n=1}^\infty \frac{1}{n^2} < \infty$$

the almost sure convergence is achieved, that is, $\widehat{\beta} \longrightarrow \beta$ almost surely.

For the limit distribution, let $Y_j = (x_j - \bar{x}_n)\varepsilon_j$. Then $\mathbb{E}(Y_j) = (x_j - \bar{x}_n)\mathbb{E}(\varepsilon_j) = 0$ and $s_n^2 = \mathbb{E}[Y_1^2 + \cdots + Y_n^2] = ((x_1 - \bar{x}_n)^2 + \cdots + (x_n - \bar{x}_n)^2)\mathbb{E}[\varepsilon_1^2] = S_{XX}\sigma_\varepsilon^2$. Assume $\max((x_1 - \bar{x}_n)^2, \ldots, (x_n - \bar{x}_n)^2)/S_{XX} \to 0$ so that the maximum variance contribution converges to zero. The Lyapounov's condition with $\delta = 2$ is

$$\frac{1}{s_n^4}\sum_{j=1}^n \mathbb{E}[Y_j^4] = \frac{1}{S_{XX}^2\sigma_\varepsilon^2}\sum_{j=1}^n (x_j - \bar{x}_n)^4\mathbb{E}[\varepsilon_j^4] \le \frac{\mathbb{E}[\varepsilon_1^4]\max((x_j - \bar{x}_n)^2)}{S_{XX}^2\sigma_\varepsilon^2}\sum_{j=1}^n (x_j - \bar{x}_n)^2 \le \frac{\mathbb{E}[\varepsilon_1^4]}{\sigma_\varepsilon^4}\frac{\max((x_j - \bar{x}_n)^2)}{S_{XX}} \to 0.$$

The central limit theorem implies

$$\frac{\bar{Y}_n}{s_n} = \frac{1}{s_n}\sum_{j=1}^n (x_j - \bar{x}_n)\varepsilon_j \xrightarrow{d} N(0,1).$$

Considering $S_{XX}/n \to \tau^2$ and $\widehat{\beta} - \beta = \bar{Y}_n/S_{XX}$,

$$\sqrt{n}(\widehat{\beta} - \beta) = \sqrt{\frac{n\sigma_\varepsilon^2}{S_{XX}}}\frac{\bar{Y}_n}{s_n} \xrightarrow{d} \frac{\sigma_\varepsilon}{\tau}N(0,1) \sim N\Big(0, \frac{\sigma_\varepsilon^2}{\tau^2}\Big).$$

Exercise 47. Show that parameter estimators in ANOVA model converge to the true parameter value in probability under a certain condition.

Autoregressive Time Series

Let ε_n be an i.i.d. sample with mean zero and finite fourth moment, that is, $\mathbb{E}(\varepsilon_n) = 0$ and $\mathbb{E}(\varepsilon_n^4) < \infty$.

Let $X_0 = 0$ and $X_n = \phi X_{n-1} + \varepsilon_n$. Then

$$\begin{aligned}X_n &= \phi X_{n-1} + \varepsilon_n = \phi(\phi X_{n-2} + \varepsilon_{n-1}) + \varepsilon_n = \phi^2 X_{n-2} + \varepsilon_n + \phi\varepsilon_{n-1} = \cdots = \phi^n X_0 + \varepsilon_n + \phi\varepsilon_{n-1} + \cdots + \phi^{n-1}\varepsilon_1 \\ &= \sum_{k=1}^n \phi^{n-k}\varepsilon_k.\end{aligned}$$

Hence mean and variance of each variable is

$$\mathbb{E}(X_n) = \sum_{k=1}^n \phi^{n-k}\mathbb{E}[\varepsilon_k] = 0$$

and

$$\mathbb{V}\text{ar}(X_n) = \mathbb{E}[X_n^2] = \mathbb{E}[\sum_{k=1}^n \phi^{n-k}\varepsilon_k \sum_{l=1}^n \phi^{n-l}\varepsilon_l] = \sum_{k=1}^n \phi^{2(n-k)}\mathbb{E}[\varepsilon_k^2] = \mathbb{E}[\varepsilon_1^2]\frac{1-\phi^{2n}}{1-\phi^2} \le \frac{\sigma_\varepsilon^2}{1-\phi^2}.$$

Let $\hat{\phi} = \underset{\phi}{argmin} \sum_{k=1}^n (X_k - \phi X_{k-1})^2$. Note that

$$\sum_{k=1}^n (X_k - \phi X_{k-1})^2 = \phi^2 \sum_{k=1}^n X_{k-1}^2 - 2\phi \sum_{k=1}^n X_k X_{k-1} + \sum_{k=1}^n X_k^2$$

is minimized at $\hat{\phi} = \sum_{k=1}^n X_k X_{k-1} / \sum_{k=1}^n X_{k-1}^2$. Then it can be expressed through ε_i's, i.e.,

$$\hat{\phi} = \frac{\sum_{k=1}^n X_{k-1} X_k}{\sum_{k=1}^n X_{k-1}^2} = \frac{\sum_{k=1}^n X_{k-1}(\phi X_{k-1} + \varepsilon_k)}{\sum_{k=1}^n X_{k-1}^2} = \phi + \frac{\sum_{k=1}^n X_{k-1}\varepsilon_k}{\sum_{k=1}^n X_{k-1}^2}$$

Let $Y_n = n^{-1} \sum_{k=1}^n X_{k-1}\varepsilon_k$. Note that X_i and ε_j are independent if $i < j$. Then

$$\begin{aligned}
\mathbb{E}(Y_n) &= \frac{1}{n}\sum_{k=1}^n \mathbb{E}(X_{k-1}\varepsilon_k) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}(X_{k-1})\mathbb{E}(\varepsilon_k) = 0, \\
\mathbb{E}(Y_n^2) &= \mathbb{E}\Big[\frac{1}{n}\sum_{k=1}^n X_{k-1}\varepsilon_k \frac{1}{n}\sum_{j=1}^n X_{j-1}\varepsilon_j\Big] = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}(X_{k-1}^2\varepsilon_k^2) + \frac{1}{n^2}\sum_{k\neq j}\mathbb{E}(X_{k-1}X_{j-1}\varepsilon_{\min(k,j)})\mathbb{E}(\varepsilon_{\max(k,j)}) \\
&= \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}(X_{k-1}^2)\mathbb{E}(\varepsilon_k^2) = \frac{1}{n}\sup_k \mathbb{E}(X_k^2)\mathbb{E}(\varepsilon_1^2) \leq \frac{\sigma_\varepsilon^4}{n(1-\phi^2)},
\end{aligned}$$

Similarly

$$\mathbb{E}(Y_n^4) = \frac{1}{n^4}\sum_{k=1}^n \mathbb{E}(X_{k-1}^4)\mathbb{E}(\varepsilon_k^4) + \frac{1}{n^4}\sum_{k\neq j} 3\mathbb{E}(X_{k-1}^2 X_{j-1}^2 \varepsilon_k^2 \varepsilon_j^2) \approx c_4/n^2$$

for some positive constant c_4. Then

$$P(|Y_n| > \epsilon) \leq \mathbb{E}(Y_n^2)/\epsilon^2 = \epsilon^{-2}\sup_k \mathbb{E}(X_k^2)\mathbb{E}(\varepsilon_1^2)/n \to 0.$$

implies $Y_n \to 0$ in probability. Also

$$\sum_{n=1}^\infty P(|Y_n| > \epsilon) \leq \mathbb{E}(Y_n^4)/\epsilon^4 \approx \epsilon^{-4} c_4 \sum_{n=1}^\infty 1/n^2 < \infty$$

for any $\epsilon > 0$ implies $Y_n \to 0$ almost surely.

It can be shown that $Z_n = n^{-1}\sum_{k=1}^n X_{k-1}^2 \to \sigma_\varepsilon^2/(1-\phi^2)$ in probability. Hence $\hat{\phi} = \phi + Y_n/Z_n \xrightarrow{p} \phi + 0/(\sigma_\varepsilon^2/(1-\phi^2)) = \phi$.

Note. Using martingale central limit theorem, $\sqrt{n}Y_n$ converges to a normal distribution.

Poisson Process

A *stochastic process* is a collection of time indexed random variables. For example, N_t is the number of clients entered a shop between time 0 and t; S_t is the price of a stock at time t.

Definition 43. A *Poisson process with intensity* λ is a stochastic process $N = \{N_t : t \geq 0\}$ taking values in non-negative integers satisfying

(a) $N_0 = 0$ and $N_s \leq N_t$ if $0 \leq s \leq t$.

(b) $P(N_{t+h} = n + m \,|\, N_t = n) = \begin{cases} 1 - \lambda h + o(h) & \text{if } m = 0, \\ \lambda h + o(h) & \text{if } m = 1, \\ o(h) & \text{if } m > 1, \end{cases}$

(c) For $0 \leq s < t$, the arrivals $N_t - N_s$ in the interval $(s, t]$ is independent of the arrivals N_s in the interval $(0, s]$.

The first assumption guarantees the process is non-decreasing. The last assumption indicates that new arrivals in $(s, t]$ does not depend on previous arrivals. The second assumption can be interpreted as no two clients arrives within very short time period and arrival rate is roughly proportional to the time length.

Theorem 96. For any fixed time $t > 0$, $N_t \sim \text{Poisson}(\lambda t)$.

Proof. Let $p_j(t) = e^{\lambda t} P(N_t = j)$. From (b) and (c),

$$p_0(t+h) = e^{\lambda(t+h)} P(N_{t+h} = 0) = e^{\lambda(t+h)} P(N_{t+h} - N_t = 0 \,|\, N_t = 0) P(N_t = 0) = e^{\lambda h}(1 - \lambda h + o(h)) p_0(t)$$

Then the derivative of $p_0(t)$ can be obtained from

$$\begin{aligned} p_0'(t) &= \lim_{h \to 0} \frac{p_0(t+h) - p_0(t)}{h} = p_0(t) \lim_{h \to 0} \frac{e^{\lambda h}(1 - \lambda h + o(h)) - 1}{h} = p_0(t) \lim_{h \to 0} [e^{\lambda h}(-\lambda + o(1)) + \frac{e^{\lambda h} - 1}{h}] \\ &= p_0(t)[-\lambda + \lambda] = 0. \end{aligned}$$

Thus $p_0(t)$ is constant and $p_0(0) = e^{\lambda \cdot 0} P(N_0 = 0) = 1$ solves $p_0(t) = 1$ and $P(N_t = 0) = e^{-\lambda t}$. For $p_1(t)$, consider

$$\begin{aligned} P(N_{t+h} = 1) &= P(N_{t+h} - N_t = 0 \,|\, N_t = 1) P(N_t = 1) + P(N_{t+h} - N_t = 1 \,|\, N_t = 0) P(N_t = 0) \\ &= e^{-\lambda h} P(N_t = 1) + (\lambda h + o(h)) e^{-\lambda t}. \end{aligned}$$

Hence $p_1(t+h) = e^{\lambda(t+h)}P(N_{t+h}=1) = e^{\lambda t}P(N_t=1) + e^{\lambda h}(\lambda h + o(h)) = p_1(t) + e^{\lambda h}(\lambda h + o(h))$ implies

$$p_1'(t) = \lim_{h\to 0}\frac{p_1(t+h)-p_1(t)}{h} = \lim_{h\to 0}\frac{e^{\lambda h}(\lambda h + o(h))}{h} = \lim_{h\to 0} e^{\lambda h}(\lambda + o(1)) = \lambda.$$

Hence $p_1(t) = \lambda t$ or $P(N_t = 1) = e^{-\lambda t}p_1(t) = \lambda t e^{-\lambda t}$.

Now assume $p_k(t) = (\lambda t)^k/k!$ for $k = 0, 1, \ldots, j-1$ for $j \geq 2$. Analogously,

$$\begin{aligned} P(N_{t+h}=j) &= P(N_{t+h}-N_t = 0 \,|\, N_t = j)P(N_t=j) + P(N_{t+h}-N_t = 1 \,|\, N_t = j-1)P(N_t = j-1) \\ &= e^{-\lambda h}P(N_t = j) + \lambda h e^{-\lambda h}\frac{(\lambda t)^{j-1}e^{-\lambda t}}{(j-1)!} + \sum_{k=2}^{j} P(N_{t+h}-N_t = k \,|\, N_t = j-k)P(N_t = j-k). \end{aligned}$$

Note that $\sum_{k=2}^{j} P(N_{t+h}-N_t = k \,|\, N_t = j-k)P(N_t = j-k) \leq P(N_{t+h}-N_t > 1 \,|\, N_t = 0) = o(h)$. Hence the derivative of $p_j(t)$ is

$$p_j'(t) = \lim_{h\to 0}\frac{p_j(t+h)-p_j(t)}{h} = \lim_{h\to 0}\frac{\lambda h(\lambda t)^{j-1}/(j-1)! + o(h)e^{\lambda(t+h)}}{h} = \frac{\lambda^j t^{j-1}}{(j-1)!}$$

and $p_j(t) = \lambda^j t^j/j!$ which implies $P(N_t = j) = e^{-\lambda t}p_j(t) = (\lambda t)^j e^{-\lambda t}/j!$. From the mathematical induction $P(N_t = j) = (\lambda t)^j e^{-\lambda t}/j! \sim \text{Poisson}(\lambda t)$. □

Let $T_0 = 0$ and $T_n = \inf\{t \geq 0 : N_t = n\}$. The *interarrival times* are the differences $X_n = T_n - T_{n-1}$ or $T_n = X_1 + \cdots + X_n$. It is easy to derive that $N_t = \max\{n \geq 0 : T_n \leq t\}$.

Theorem 97. The interarrival times $X_1, X_2, \ldots$ are independent and identically distributed from exponential with λ.

Proof. Note $P(X_1 > t) = P(N_t = 0) = e^{-\lambda t}$ implies $X_1 \sim \text{exponential}(\lambda)$. Also $P(X_2 > t \,|\, X_1 = t_1) = P(N_{t_1+t} - N_{t_1} = 0 \,|\, N_{t_1} = 1) = e^{-\lambda t}$. Similarly $P(X_{n+1} > t \,|\, X_1 = t_1, \ldots, X_n = t_n) = P(N_{t_1+\cdots+t_n+t} - N_{t_1+\cdots+t_n} = 0 \,|\, N_{t_1+\cdots+t_n} = n) = e^{-\lambda t}$. Hence $X_1, \ldots, X_n$ are identically distributed and independent. □

It is known that if $X \sim \text{Poisson}(\mu)$ and $Y \sim \text{Poisson}(\lambda 0$ are independent, then $X + Y \sim \text{Poisson}(\mu+\lambda)$ also $X \,|\, X+Y = n \sim \text{binomial}(n, \mu/(\mu+\lambda))$.

Example 105. The number of bus arrived at a bus stop follows a Poisson process with parameter $\lambda = 10$. When 8 buses arrived between time 0 and 1, what is the probability that no buses arrived between time 0.5 and 1?

Let $X = N_{0.5} \sim \text{Poisson}(10\times 0.5) \sim \text{Poisson}(5)$ and $Y = N_1 - N_{0.5} \sim \text{Poisson}(10\times(1-0.5)) \sim \text{Poisson}(5)$. Then X and Y are independent and $X + Y = N_1$. So $P(N_{0.5} = 0 \,|\, N_1 = 8) = P(X = 0 \,|\, X+Y = 8) \sim$

$P(Z=0) = \binom{8}{0}(1/2)^0(1/2)^8 = 1/256 = 0.0039.$

If the color of buses are red and purple with probability 0.4 and 0.6, then what is the expectation of red cars arriving between time 0 and 2?

Let R be the number of red cars arriving between time 0 and 2. Then $N_2 \sim$ Poisson(20) and

$$P(R=k) = \sum_{n=0}^{\infty} P(R=k \mid N_2=n)P(N_2=n) = \sum_{n=k}^{\infty}\binom{n}{k}(0.4)^k(0.6)^{n-k} \times e^{-20}20^n/n! = \frac{(0.4)^k 20^k e^{-20}}{k!}\sum_{n=k}^{\infty} 12^{n-k}/(n-k)! =$$

Finally $\mathbb{E}[R] = \mathbb{E}[\text{Poisson}(8)] = 8$.

As time goes, the average number of buses arriving the bus stop per unit time converges to 10 almost surely because

$$\frac{N_t}{N_t+1}\frac{1}{\bar{X}_{N_t+1}} = \frac{N_t}{X_1+\cdots+X_{N_t+1}} \le \frac{N_t}{t} \le \frac{N_t}{X_1+\cdots+X_{N_t}} = \frac{1}{\bar{X}_{N_t}}$$

converges to $1/\mathbb{E}(X_1) = 1/(1/\lambda) = \lambda$ almost surely by the strong law of large numbers.

Exercise 48. Let N_t be a Poisson process with parameter λ. Show that $N_t/t \to \lambda$ almost surely as $t \to \infty$.

Example 106 (Poisson approximation)**.** The average number of earthquake is reported μ times per year. Suppose the occurrence of earthquake does not depend on previous occurrence. Q: What is the probability of having k earthquakes in a year?

The time range (a year) can be split into n equal length periods. Take $n > \mu$ big enough to have there is no more than 1 earthquake per each period. Then the expected probability of having an earthquake per period is $p_n = \mu/n$ and the total number of earthquake happened within the year becomes a binomial distribution with parameters n and p_n, that is, $Y_n \sim \text{binomial}(n, p_n)$. Hence the probability having k earthquakes is

$$\begin{aligned}\lim_{n\to\infty} P(Y_n=k) &= \lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k} = \lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\mu}{n}\right)^k\left(1-\frac{\mu}{n}\right)^{n-k} \\ &= \frac{\mu^k}{k!}\lim_{n\to\infty} e^{-(\mu/n)(n-k)} = \frac{\mu^k}{k!}e^{-k}.\end{aligned}$$

Hence the number of earthquakes per year follows a Poisson distribution.