Matematika Statistika

Statistik dan Distribusi Sampling

Raden Aurelius Andhika Viadinugroho

Institut Teknologi Sepuluh Nopember
raden.aureliusav@its.ac.id

7 Agustus 2024

Daftar Isi

Populasi $X \sim f(x|\theta)$, bentuk f diketahui tetapi mengandung parameter $\theta \in \Omega$.

Contoh 1.1

1. $X \sim f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x > 0$
   $\Omega\{\theta : \theta > 0\}$
2. $X \sim p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, x = 0, 1, 2, \ldots$
   $\Omega\{\theta : 0 < \theta < 1\}$
3. $X \sim f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, x \geq 0$
   $\Omega\{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

Keterangan: Ω disebut ruang parameter.

Tujuan analisis adalah menentukan nilai θ yang paling masuk akal (estimasi titik) atau menentukan himpunan bagian dari untuk menyatakan bahwa daerah tersebut memuat θ atau tidak (uji hipotesis). Analisa didasarkan pengamatan sampel random.

Definisi 1.2

Variabel random $X_1, X_2, \ldots, X_n$ membangun sampel random pada variabel random X jika $X_1, X_2, \ldots, X_n$ i.i.d.

Definisi 1.3

Misalkan n variabel random $(X_1, X_2, \ldots, X_n)$ merupakan sampel random dari variabel random X. Maka sebarang fungsi $T = T(X_1, X_2, \ldots, X_n)$ dari sampel disebut statistik.

- Karena statistik adalah fungsi sampel, maka statistik juga merupakan variabel random. Statistik juga sering digunakan untuk meringkas data.
- Suatu statistik $T = T(X_1, X_2, \ldots, X_n)$ bisa memberi informasi tentang parameter θ yang tidak diketahui.
- Dalam kasus seperti itu, statistik yang bersesuaian disebut estimator titik dari θ.
- Bila $E(T) = \theta$, maka T disebut estimator takbias dari θ, dan bila $T_n \xrightarrow{p} \theta$, T disebut estimator yang konsisten dari θ.

Contoh 1.4

Pandang persoalan kotak dengan m bola. Dimisalkan bola-bolanya identik kecuali nomornya dan misalkan m tidak diketahui. Untuk mengestimasi m diambil sampel random $(X_1, X_2, \ldots, X_n)$ dengan pengembalian. Distribusi dari masing-masing X_i adalah $P(X = x) = \dfrac{1}{m}$ untuk $x = 1, 2, \ldots, m$. Secara intuitif, estimator titik dari m adalah statistik $T = \max(X_1, X_2, \ldots, X_n)$.

- Pertanyaannya, seberapa 'dekat' T dengan m agar bisa disebut estimator yang 'baik'?
- Untuk menjawab pertanyaan ini kita harus mencari distribusi dari T. Support dari T adalah $\{1, 2, \ldots, m\}$.
- Untuk menentukan fungsi distribusi kumulatif dari T, karena T adalah maksimum observasi X, maka kejadian $T \leq t$ untuk $1 \leq t \leq m$ adalah:

$$\{T \leq t\} = \{X_1 \leq t, \ldots, X_n \leq t\} = \bigcap_{i=1}^{n} \{X_i \leq t\}$$

Menggunakan kenyataan bahwa $(X_1, X_2, \ldots, X_n)$ i.i.d., maka fungsi distribusi kumulatif T adalah

$$P(T \leq t) = \prod_{i=1}^{n} P(X_i \leq t) = P(X_i \leq t)^n = (\frac{\lfloor t \rfloor}{m})^n$$

dengan $\lfloor t \rfloor$ menyatakan bilangan bulat terbesar yang lebih kecil atau sama dengan t. Sehingga untuk $0 \leq t \leq m$,

$$P(T_n \leq t) = (\frac{\lfloor t \rfloor}{m})^n \rightarrow \begin{cases} 0 & \text{jika } t < m \\ 1 & \text{jika } t = m \end{cases}$$

- Berdasarkan hasil tersebut, artinya $T_n \xrightarrow{d} m$.
- Menggunakan sifat konvergensi ke suatu konstanta, karena $T_n \xrightarrow{d} m$, maka $T_n \xrightarrow{p} m$. Jadi T_n estimator konsisten untuk m.
- Perhatikan untuk kasus $E(X) = \dfrac{m+1}{2}$. Akibatnya $E(2\bar{X} - 1) = m$, dengan $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$. Artinya, $2\bar{X} - 1$ adalah estimator takbias untuk m.

- Misalkan $(X_1, X_2, \ldots, X_n)$ merupakan sampel random dari variabel random X.
- Statistik $T(X_1, X_2, \ldots, X_n)$ adalah variabel random dengan distribusi yang pada dasarnya dapat dijabarkan dari distribusi gabungan $X_1, X_2, \ldots, X_n$.
- Distribusi T sering disebut distribusi sampling statistik T karena dijabarkan melalui distribusi gabungan sampel terobservasi.
- Dalam bab ini akan dijabarkan distribusi sampling yang terkait dengan distribusi normal karena hasilnya akan membawa pada distribusi-distribusi yang memainkan peran penting dalam statistika inferensi.

Teorema 2.1

Jika $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \ldots, n$ *menyatakan* n *variabel random normal independen, maka:*

$$Y = \sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$$

Bukti.

Misalkan $M_Y(t) = E(e^{ty})$ adalah MGF Y, maka

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(a_i t) \\ &= \prod_{i=1}^{n} e^{a_i \mu_i t + \frac{1}{2} a_i^2 \sigma_i^2 t^2} \\ &= \exp(t \sum_{i=1}^{n} a_i \mu_i + t^2 \sum_{i=1}^{n} \frac{a_i^2 \sigma_i^2}{2}) \end{aligned}$$

yang merupakan MGF variabel random normal dengan mean

$Y \sim N(\sum^{n} a_i \mu_i, \sum^{n} a_i^2 \sigma_i^2)$ $\square$

Akibat 2.2

Jika $X_1, X_2, \ldots, X_n$ menyatakan sampel random dari $N(\mu, \sigma^2)$, maka $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Beberapa sifat distribusi sampling berhubungan erat dengan distribusi gamma khusus yang disebut dengan distribusi khi-kuadrat (*chi-square*).

Definisi 2.3

Variabel random X disebut berdistribusi gamma dengan parameter α dan β - dinotasikan dengan $X \sim GAM(\alpha, \beta)$ - jika

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}; x > 0, \alpha > 0, \beta > 0$$

Teorema 2.4

Jika $X \sim GAM(\alpha, \beta)$*, maka* $M_X(t) = \dfrac{1}{(1-\beta t)^{\alpha}}$, $E(X) = \alpha\beta$*, dan* $Var(X) = \alpha\beta^2$.

Definisi 2.5

Jika $X \sim GAM(\alpha, \beta)$ *dengan* $\alpha = \dfrac{n}{2}$ *dan* $\beta = 2$*, maka variabel random* X *disebut berdistribusi khi-kuadrat dengan derajat kebebasan n ditulis* $\chi^2_{(n)}$*. Artinya, jika* $X \sim \chi^2_{(n)}$ *maka* $E(X) = n$, $Var(X) = 2n$*, dan* $M_X(t) = \dfrac{1}{(1-2t)^{\frac{n}{2}}}$

Teorema 2.6

Jika $X \sim GAM(\alpha, \beta)$, *maka* $Y = \dfrac{2X}{\beta} \sim \chi^2(2\alpha)$.

Bukti.

$$
\begin{aligned}
M_Y(t) &= M_{\frac{2X}{\beta}}(t) \\
&= M_X(\frac{2t}{\beta}) \\
&= (1 - 2t)^{-\frac{2\alpha}{2}}
\end{aligned}
$$

yang merupakan fungsi pembangkit momen distribusi khi-kuadrat dengan derajat kebebasan sama dengan 2α. $\square$

Teorema berikut menunjukkan sifat penting yang menyatakan bahwa jumlahan variabel random berdistribusi khi-kuadrat yang saling independen juga berdistribusi khi-kuadrat.

Teorema 2.7

Jika $X_i \sim \chi^2(n_i), i = 1, 2, \ldots, m$ *merupakan variabel random khi-kuadrat yang saling independen maka*

$$V = \sum_{i=1}^{m} X_i \sim \chi^2(\sum_{i=1}^{m} n_i)$$

Bukti.

$$
\begin{aligned}
M_V(t) &= (1-2t)^{-\frac{n_1}{2}}(1-2t)^{-\frac{n_2}{2}}\dots(1-2t)^{-\frac{n_m}{2}} \\
&= (1-2t)^{-\frac{\sum_{i=1}^{m} n_i}{2}}
\end{aligned}
$$

yang merupakan fungsi pembangkit momen distribusi khi-kuadrat dengan derajat kebebasan $\displaystyle\sum_{i=1}^{m} n_i$. □

Teorema 2.8

Jika $Z \sim N(0,1)$, *maka* $Z^2 \sim \chi^2(1)$.

Bukti.

$$
\begin{aligned}
M_{Z^2}(t) &= E(e^{tz^2}) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz^2 - \frac{z^2}{2}}\, dz \\
&= \frac{1}{(1-2t)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \frac{(1-2t)^{\frac{1}{2}}\sqrt{2\pi}}{e}^{-\frac{z^2}{2}(1-2t)}\, dz \\
&= \frac{1}{(1-2t)^{\frac{1}{2}}}
\end{aligned}
$$

yang merupakan fungsi pembangkit momen distribusi khi-kuadrat dengan derajat kebebasan satu. □

Akibat 2.9

Jika $X_1, X_2, \ldots, X_n$ menyatakan sampel random dari $N(\mu, \sigma^2)$, maka

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)}$$

$$\frac{n(\bar{x} - \mu)^2}{\sigma^2} \sim \chi^2_{(1)}$$

Dari akibat diatas, telah diketahui jika $X_1, X_2, \ldots, X_n$ menyatakan sampel random dari $N(\mu, \sigma^2)$, maka $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Perhatian berikutnya adalah distribusi dari variansi sampel $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$. Distribusi dari s^2 tidak bisa mengikuti secara langsung dari akibat diatas karena $X_i - \bar{X}$ tidak independen. Secara fungsional bentuk tersebut dependen karena $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Teorema 2.10

Jika $X_1, X_2, \ldots, X_n$ menyatakan sampel random dari $N(\mu, \sigma^2)$, maka

1. $\bar{x}$ *dan suku-suku* $x_i - \bar{x}, i = 1, 2, \ldots, n$ *independen.*
2. $\bar{x}$ *dan* s^2 *independen.*
3. $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

Teorema 2.11

Jika Z variabel random $N(0,1)$, dan U variabel random $\chi^2_{(r)}$, Z dan U independen, maka

$$T = \frac{Z}{\sqrt{\frac{U}{r}}}$$

mempunyai distribusi t dengan derajat kebebasan r, dengan PDF

$$f(t) = \frac{\Gamma[\frac{(r+1)}{2}]}{\sqrt{\pi r}\Gamma(\frac{r}{2})(1+\frac{t^2}{r})^{\frac{(r+1)}{2}}}$$

Teorema 2.12

Jika $X_1, X_2, \ldots, X_n$ menyatakan sampel random dari $N(\mu, \sigma^2)$, maka

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

Bukti.

Karena $z = \sqrt{n}(\frac{\bar{x} - \mu}{\sigma}) \sim N(0, 1)$, dari teorema di atas diperoleh $v = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Selanjutnya, $\bar{x}$ dan s independen sehingga $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}}$ menurut teorema di atas

mempunyai distribusi t dengan derajat kebebasan $n - 1$. $\square$

Teorema 2.13

Jika U dan V variabel random independen yang masing-masing berdistribusi khi-kuadrat dengan derajat kebebasan r_1 dan r_2, maka

$$F = \frac{\frac{U}{r_1}}{\frac{V}{r_2}}$$

mempunyai distribusi F dengan derajat kebebasan r_1 dan r_2. PDF yang bersesuaian adalah

$$f(w) = \frac{\Gamma[\frac{(r_1+r_2)}{2}](\frac{r_1}{r_2})^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1}}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})(1+\frac{r_1 w}{r_2})^{\frac{(r_1+r_2)}{2}}}, 0 < w < \infty$$

Teorema 2.14

Jika X variabel random berdistribusi F, yaitu $X \sim F(r_1, r_2)$, maka variabel random $Y = \dfrac{(\frac{r_1}{r_2})X}{1 + (\frac{r_1}{r_2})X}$ mempunyai PDF

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\; 0 < y < 1$$

dengan $\alpha = \dfrac{r_1}{2}$ dan $\beta = \dfrac{r_2}{2}$. Selanjutnya, Y dikatakan berdistirbusi Beta dengan parameter $\alpha > 0$ dan $\beta > 0$ - atau dapat dinotasikan dengan $Y \sim Beta(\alpha, \beta)$ - dengan $E(Y) = \dfrac{\alpha}{\alpha+\beta}$ dan $Var(Y) = \dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.