Lecture 14: Sampling from Normal Distribution

Mathematical Statistics I, MATH 60061/70061

Tuesday October 26, 2021

Reference: Casella & Berger, 5.3

Recap: PDF of a location-scale transformation

Let X have PDF f_X, and let $Y = a + bX$, with $b \neq 0$. Let $y = a + bx$, to mirror the relationship between Y and X. Then $\frac{dy}{dx} = b$, so the PDF of Y is

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X\left(\frac{y-a}{b}\right) \frac{1}{|b|}.$$

Change of variables in multiple dimensions

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a continuous random vector with joint PDF $f_{\boldsymbol{X}}$. Let $\boldsymbol{Y} = g(\boldsymbol{X})$, and mirror this by letting $\boldsymbol{y} = g(\boldsymbol{x})$. Suppose g is invertible, so we have $\boldsymbol{X} = g^{-1}(\boldsymbol{Y})$ and $\boldsymbol{x} = g^{-1}(\boldsymbol{y})$.

Suppose that all partial derivatives $\partial x_i / \partial y_j$ exist and are continuous, so we can form the **Jacobian matrix**

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Also assume that the determinant of this Jacobian matrix is never 0. Then the joint PDF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(\boldsymbol{g}^{1}(\boldsymbol{y})\right) \cdot \left| \left| \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} \right| \right|.$$

Recap: Gamma distribution

The Gamma distribution is a continuous distribution on the positive real line, which generalizes the Exponential distribution.

A random variable Y is said to have the **Gamma distribution** with parameters a and λ, $Y \sim \mathrm{Gamma}(a, \lambda)$, where $a > 0$ and $\lambda > 0$, if its PDF is

$$f(y) = \frac{1}{\Gamma(a)}(\lambda y)^a e^{-\lambda y}\frac{1}{y}, \quad y > 0.$$

where Γ is the **gamma function**, defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x}\frac{dx}{x},$$

for real numbers $a > 0$.

Mean and variance of $\mathrm{Gamma}(a, \lambda)$

Mean of $X \sim \mathrm{Gamma}(a, 1)$:

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot \frac{1}{\Gamma(a)} x^a e^{-x} \frac{dx}{x} = \frac{1}{\Gamma(a)} \int_0^\infty x^{a+1} e^{-x} \frac{dx}{x} \\ &= \frac{\Gamma(a+1)}{\Gamma(a)} = a. \end{aligned}$$

Variance of $X \sim \mathrm{Gamma}(a, 1)$:

$$E(X^2) = \int_0^\infty \frac{1}{\Gamma(a)} x^{a+2} e^{-x} \frac{dx}{x} = \frac{\Gamma(a+2)}{\Gamma(a)} = (a+1)a,$$

$$\mathrm{Var}(X) = (a+1)a - a^2 = a.$$

For $Y = X/\lambda \sim \mathrm{Gamma}(a, \lambda)$,

$$E(Y) = \frac{1}{\lambda} E(X) = \frac{a}{\lambda}, \qquad \mathrm{Var}(Y) = \frac{1}{\lambda^2} \mathrm{Var}(Y) = \frac{a}{\lambda^2}.$$

A sum of independent Gamma RVs

Let $X_1, \ldots, X_n$ be independent with $X_j \sim \mathrm{Gamma}(a_j, \lambda)$. What is the distribution of $X_1 + \cdots + X_n$?

The $\mathrm{Gamma}(a_j, \lambda)$ MGF is $\left(\frac{\lambda}{\lambda - t}\right)^{a_j}$ for $t < \lambda$, so the MGF of $X_1 + \cdots + X_n$ is

$$M_n(t) = \left(\frac{\lambda}{\lambda - t}\right)^{(a_1 + \cdots + a_n)}, \quad \text{for } t < \lambda.$$

This is the MGF of $\mathrm{Gamma}(\sum_{i=1}^{n} a_i, \lambda)$

Chi-Squared distribution

Let $V = Z_1^2 + \cdots + Z_n^2$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0, 1)$. Then V is said to have the **Chi-Squared distribution with** n **degrees of freedom**. We write this as $V \sim \chi_n^2$.

Chi-Squared distribution

Let $V = Z_1^2 + \cdots + Z_n^2$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$. Then V is said to have the **Chi-Squared distribution with** n **degrees of freedom**. We write this as $V \sim \chi_n^2$.

The PDF of $Z_1^2 \sim \chi_1^2$ equals the PDF of the $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$: for $x > 0$,

$$F(x) = P(Z_1^2 \leq x) = P(-\sqrt{x} \leq Z_1 \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1,$$

so

$$f(x) = \frac{d}{dx} F(x) = 2\varphi(\sqrt{x}) \frac{1}{2} x^{-1/2} = \frac{1}{\sqrt{2\pi x}} e^{-x/2},$$

which is the $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$ PDF.

Chi-Squared distribution

Let $V = Z_1^2 + \cdots + Z_n^2$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$. Then V is said to have the **Chi-Squared distribution with n degrees of freedom**. We write this as $V \sim \chi_n^2$.

The PDF of $Z_1^2 \sim \chi_1^2$ equals the PDF of the $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$: for $x > 0$,

$$F(x) = P(Z_1^2 \leq x) = P(-\sqrt{x} \leq Z_1 \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1,$$

so

$$f(x) = \frac{d}{dx}F(x) = 2\varphi(\sqrt{x})\frac{1}{2}x^{-1/2} = \frac{1}{\sqrt{2\pi x}}e^{-x/2},$$

which is the $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$ PDF.

Then, because $V = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$ is the sum of n independent $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$ random variables, we have $V \sim \mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$.

Mean and variance of χ_n^2

A χ_n^2 random variable V has the $\mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$ distribution. So, from our knowledge about the Gamma, we have

$$E(V) = \frac{n/2}{1/2} = n, \quad \mathrm{Var}(V) = \frac{n/2}{(1/2)^2} = 2n.$$

Plugging $n/2$ and $1/2$ into the more general $\mathrm{Gamma}(a, \lambda)$ MGF gives the MGF of the Chi-Square distribution:

$$M_V(t) = \left(\frac{1/2}{1/2 - t}\right)^{n/2} = \left(\frac{1}{1 - 2t}\right)^{n/2}.$$

Chi-Squared is a special case of the Gamma distribution

The Chi-Squared PDF with n degrees of freedom ($\mathrm{Gamma}(n/2, 1/2)$ PDF) is given by

$$f(x) = \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} x^{p/2-1} e^{-x/2}, \quad 0 < x < \infty.$$

1. If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_1$.
2. If $X_1, \ldots, X_n$ are independent and $X_i \sim \chi^2_{p_i}$, then $X_1 + \cdots + X_n \sim \chi^2_{p_1+\cdots+p_n}$.

Normal sample mean and variance

Let $X_1, ..., X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ and let $\bar{X}$ and S^2 be the sample mean and sample variance. Then

1. $\bar{X}$ and S^2 are independent random variables.
2. $\bar{X}$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution.
3. $(n-1)S^2/\sigma^2$ has a Chi-Squared distribution with $n-1$ degrees of freedom.

Independence between Normal sample mean and variance

Without loss of generality, assume $\mu = 0$ and $\sigma = 1$.

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1}\left(\left[-\sum_{i=2}^{n}(X_i - \bar{X})\right]^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1}\left(\left[\sum_{i=2}^{n}(X_i - \bar{X})\right]^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right).
\end{aligned}$$

Thus, S^2 can be written as a function only of
$(X_2 - \bar{X}, \ldots, X_n - \bar{X})$

Independence between Normal sample mean and variance

The joint PDF of the sample $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^n x_i^2}, \quad -\infty < x_i < \infty.$$

Consider a linear transformation:

$$\begin{aligned} y_1 &= \bar{x} \\ y_2 &= x_2 - \bar{x} \\ &\vdots \\ y_n &= x_n - \bar{x}, \end{aligned}$$

where $\boldsymbol{g}^{-1}(\boldsymbol{y}) = (y_1 - \sum_{i=2}^n y_i, y_2 + y_1, \ldots, y_n + y_1)$, and the Jacobian equal to $1/n$.

Independence between Normal sample mean and variance

The joint PDF of $Y_1, \ldots, Y_n$ is

$$
\begin{aligned}
&f(y_1, \ldots, y_n) \\
&\quad = \frac{n}{(2\pi)^{n/2}} e^{-(1/2)(y_1 - \sum_{i=2}^{n} y_i)^2} e^{-(1/2)\sum_{i=2}^{n}(y_i + y_1)^2} \\
&\quad = \left[\left(\frac{n}{2\pi} \right)^{1/2} e^{-n y_1^2/2} \right] \left[\frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-(1/2)[\sum_{i=2}^{n} y_i^2 + (\sum_{i=2}^{n} y_i)^2]} \right],
\end{aligned}
$$

for $-\infty < y_i < \infty$.

The joint PDF of $Y_1, \ldots, Y_n$ factors, so Y_1 is independent of $Y_2, \ldots, Y_n$.

Since $\bar{X} = Y_1$ and S^2 is a function of $Y_2 \ldots, Y_n$, $\bar{X}$ is independent of S^2.

Normal sample variance is a scaled Chi-Squared

Note

$$
\begin{aligned}
(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu - (\bar{X} - \mu))^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.
\end{aligned}
$$

Then

$$
n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2
$$

Since $Z_i \sim \mathcal{N}(0,1)$ and $Z_1, \ldots, Z_n$ are independent, we have shown that

- Each $Z_i^2 \sim \chi_1^2$
- The sum $\sum_{i=1}^{n} Z_i^2$ is a χ_n^2, and its MGF is $(1-2t)^{-n/2}$, $t < 1/2$
- $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0,1)$ and hence $n[(\bar{X} - \mu)/\sigma]^2 \sim \chi_1^2$

Normal sample variance is a scaled Chi-Squared

In the expression

$$n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2+\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n} Z_i^2$$

- The LHS is a sum of two independent random variables. Hence, if $M(t)$ is the MGF of $(n-1)S^2/\sigma^2$, then the MGF of the sum on LHS is

$$(1-2t)^{-1/2}M(t)$$

- Since the RHS has MGF $(1-2t)^{-n/2}$, we must have

$$M(t)=(1-2t)^{-(n-1)/2} \quad t<1/2,$$

  which is the MGF of χ^2_{n-1}.

Inference about μ with a Normal random sample

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. We know

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- If σ is known, we can use the above expression as a basis for inference about μ.
- If both μ and σ are unknown (as in most cases), we consider the quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

  as a basis for inference about μ.

t distribution

Note

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}}$$

- The numerator is a $\mathcal{N}(0,1)$ random variable.
- The denominator is a $\sqrt{\chi^2_{n-1}/(n-1)}$ random variable, independent of the numerator.

The quantity $(\bar{X}-\mu)/(S/\sqrt{n})$ is said to have t distribution with $n-1$ degrees of freedom. Equivalently, a random variable T has t distribution with p degrees of freedom, and we write $T \sim t_p$ if it has PDF is given by

$$f_T(t) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}\frac{1}{(p\pi)^{1/2}}\frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad -\infty < t < \infty.$$

F distribution (distribution of a ratio of variances)

Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu_X, \sigma_X^2)$, $Y_1, \dots, Y_m$ be a random sample from $\mathcal{N}(\mu_Y, \sigma_Y^2)$, X_i's and Y_i's be independent, and S_X^2 and S_Y^2 be the sample variances based on X_i's and Y_i's, respectively.

From the previous discussion, $(n-1)S_X^2/\sigma_X^2$ and $(m-1)S_Y^2/\sigma_Y^2$ are both Chi-Squared random variables, and the ratio

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

has the F distribution with degrees of freedom $n-1$ and $m-1$ (denoted by $F_{n-1,m-1}$).

Properties of F distribution

Let $F_{p,q}$ denote the F distribution with degrees of freedom p and q. The PDF of $F_{p,q}$ is given by

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{(p/2)-1}}{(1+(p/q)x)^{(p+q)/2}}, \quad 0 < x < \infty.$$

1. If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
2. If X has the t distribution with degrees of freedom q, then $X^2 \sim F_{1,q}$.
3. If $X \sim F_{p,q}$, then $(p/q)X/(1+(p/q)X) \sim \mathrm{Beta}(p/2, q/2)$.

The first two properties follow directly from the definitions of F and t distributions.

F distribution

Note that $Z = (p/q)X$ has PDF

$$\frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \frac{z^{p/2-1}}{(1+z)^{(p+q)/2}}, \qquad z > 0$$

Let $u = z/(1+z)$. Then $z = u/(1-u)$, $dz = (1-u)^{-2}du$, and the PDF of $U = Z/(1+Z)$ is

$$\frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{u}{1-u}\right)^{p/2-1} \frac{1}{(1-u)^{-(p+q)/2}} \frac{1}{(1-u)^2}$$

$$= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} u^{p/2-1}(1-u)^{q/2-1}, \quad u > 0$$