Lecture 16: October 11

Lecturer: Donlapark Pornnopparath

From the previous lecture, we learned that the mgf of a normal random variable $X \sim N(\mu, \sigma^2)$ is

$$M_X(t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2}. \tag{16.1}$$

With this, give a random sample from a normal distribution, we can find the distribution of the sample mean.

Theorem 16.1. Suppose that $X_1, X_2, \ldots, X_n$ are iid sample from the distribution $N(\mu, \sigma^2)$ then

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

Proof:

$$\begin{aligned}
M_{\bar{X}}(t) &= \mathbb{E}[e^{(t/n)X_1} e^{(t/n)X_2} \ldots e^{(t/n)X_n}] \\
&= \mathbb{E}[e^{(t/n)X_1}]^n && \text{(Independence)} \\
&= [M_{X_1}(t/n)]^n \\
&= \left[\exp\left(\mu \frac{t}{n} + \frac{\sigma^2 (t/n)^2}{2}\right)\right]^n \\
&= \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right).
\end{aligned}$$

Comparing the last line with (16.1) we see that $\bar{X} \sim N(\mu, \sigma^2/n)$. ■

We now study the distribution of S^2/σ^2 which we will see that is related to a so-called chi-squared distribution.

Definition 16.2. Let $Z_1, Z_2, \ldots, Z_p \sim N(0, 1)$ independent, then

$$\chi_p^2 \sim Z_1^2 + Z_2^2 + \ldots + Z_p^2 \tag{16.2}$$

is called a *chi-squared* random variable with p degrees of freedom.

The mgf of χ_1^2 is actually pretty simple.

Proposition 16.3. Suppose that $X \sim \chi_p^2$. Then

$$M_X(t) = \frac{1}{(1-2t)^{p/2}}. \tag{16.3}$$

Proof: We start from the case $p = 1$. Suppose that $X \sim \chi_1^2$. This follows directly from the definition of mgf and (16.2):

$$M_X(t) = \mathbb{E}[e^{tZ^2}] = \int e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$
$$= \int \frac{1}{2\pi} e^{-(\frac{1}{2}-t)z^2}\, dz$$

By a change of variables $u = \sqrt{1-2t}z$, we have

$$M_X(t) = \frac{1}{\sqrt{1-2t}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$
$$= \frac{1}{\sqrt{1-2t}}.$$

For the general case: $X \sim \chi^2_p$, we note that $X = X_1 + X_2 + \ldots + X_p$ where $X_1, X_2, \ldots, X_p \sim \chi^2_1$ are independent. It then follows that

$$M_X(t) = \mathbb{E}[e^{t(X_1+X_2+\ldots+X_p)}] = \left(\mathbb{E}[e^{tX_1}]\right)^p = \frac{1}{(1-2t)^{p/2}}$$

■

The following proposition is the last thing we need before studying the distribution of S^2/σ^2.

Proposition 16.4. $\overline{X}$ and S^2 are independent.

Proof: It suffices to proof that the following two random variables are independent:

$$\overline{X} \quad \text{and} \quad Y = \left(X_2 - \overline{X}, X_3 - \overline{X}, \ldots, X_n - \overline{X}\right). \tag{16.4}$$

This is because $\sum_{i=2}^n (X_i - \overline{X}) = -(X_1 - \overline{X})$ and so S^2 is a function of Y. We will compute the joint pdf of $(\overline{X}, Y)$ and show that it factors into two pdfs of $\overline{X}$ and Y. This can be done by transforming the joint pdf of $X_1, X_2, \ldots, X_n$:

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_i (x-\mu)^2}{2\sigma^2}\right\}$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_i (x-\overline{x})^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\}.$$

The middle term only depends on $y = (x_2 - \overline{x}, x_3 - \overline{x}, \ldots, x_n - \overline{x})$, while the last term only depends on $\overline{x}$. Therefore, by making a transformation:

$$(x_1, x_2, \ldots x_n) \mapsto (\overline{x}, x_2 - \overline{x}, x_3 - \overline{x}, \ldots, x_n - \overline{x}),$$

whose Jacobian matrix is a (nonzero) constant, the joint pdf of $\overline{X}$ and Y factors completely into a function of $\overline{x}$ and y. ■

We are now ready to find out the probability distribution of S^2/σ^2.

Theorem 16.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ and S^2 be the sample variance, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Proof: First, we introduce the following random variables:

$$U = \frac{(n-1)S^2}{\sigma^2}$$
$$V = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$
$$W = \frac{n(\overline{X} - \mu)^2}{\sigma^2} \sim \chi_1^2,$$

which makes $V = U + W$. From Proposition 16.4, S^2 and W are independent and it follows that

$$\begin{aligned} M_V(t) &= M_U(t) M_W(t) \\ M_U(t) &= \frac{M_W(t)}{M_V(t)} \\ &= \frac{1/(1-2t)^{1/2}}{1/(1-2t)^{n/2}} \\ &= (1-2t)^{-(n-1)/2}. \end{aligned}$$

The result follows from the uniqueness of mgf. ■

16.1 Student's t distribution

Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ be independent. From Theorem 16.1 we have that

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

Knowing the value of σ beforehand would help in estimating the value of μ. However, σ is usually unknown, preventing us from doing the usual inference tasks. To address this problem, W. S. Gosset (who had the pseudonym Student), came up with an analogous random variable,

$$\frac{\overline{X} - \mu}{S^2/\sqrt{n}}. \tag{16.5}$$

We can derive its probability distribution from what we have already learned. First, we divide both the numerator and the denominator of (16.5) by σ:

$$\frac{\overline{X} - \mu}{S^2/\sqrt{n}} = \frac{(\overline{X} - \mu)/\sigma}{S/(\sigma\sqrt{n})} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}.$$

Here, we see two familiar random variables: $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$ and $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$, which are independent from Proposition 16.4.

Definition 16.6. Let $U \sim N(0, 1)$ and $V \sim \chi_p^2$ be independent. Then we say that the random variable

$$T = \frac{U}{\sqrt{V/p}}$$

has the Student's t distribution with p degrees of freedom.

This makes a study of Student's t distribution, say T, relatively simple as the joint pdf of U and V factors:

$$f_{U,V}(u,v) = f_U(u)f_V(v)$$

Now we can make a change of variables:

$$(u,v) \mapsto \left(t = \frac{u}{\sqrt{v/(n-1)}}, w = v\right)$$

and obtain the joint pdf $f_{T,W}(t,w)$. The pdf of T can now be computed integrating with respect to w.

16.2 F distribution

Let $X_1, X_2 \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ be independent and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$ be independent. Sometimes we want to compare the variance between the populations, and so the ratio σ_X^2/σ_Y^2 would be of interest here. An intuitive approximation of this would be S_X^2/S_Y^2, and we might measure the similarity between these two quantities using

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

Here, we see two independent chi-squared random variables: $U = (n-1)S_X^2/\sigma_X^2 \sim \chi_{n-1}^2$ and $V = (n-1)S_Y^2/\sigma_Y^2 \sim \chi_{m-1}^2$. This is a special case of the F *distribution*:

Definition 16.7. Let $U \sim \chi_p^2$ and $V \sim \chi_q^2$ be independent. Then we say that the random variable

$$F = \frac{U/p}{V/q}$$

has the F distribution with p and q degrees of freedom.

in the our context, $p = n-1$ and $q = m-1$.