ST202 Lent Term – Handout 1

Yudong Chen
y.chen276@lse.ac.uk

February, 2023

1 Gamma, Chi-squared, *t*-, *F*- and Beta distributions

We have studied a lot about Gamma distributions last term. This term, we will encounter a few more new distributions, which will be useful when we discuss point estimation and hypothesis testing in the next couple of weeks.

1.1 Gamma distribution

Definition and basic properties

There are two ways to parameterise a Gamma distribution. We use throughout this course a shape parameter $\alpha > 0$ and a rate parameter $\theta > 0$ to characterise. We write the distribution as $\text{Gamma}(\alpha, \theta)$ or $\Gamma(\alpha, \theta)$. The latter shall not be confused with the Gamma function, though. The density of $X \sim \text{Gamma}(\alpha, \theta)$ is:

$$f_X(x) = \frac{\theta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x} \qquad \text{for } x > 0.$$

[The second way to parameterise a Gamma distribution is to use a scale parameter $\beta > 0$ instead. The scale parameter is simply the reciprocal of the rate parameter $\beta = 1/\theta$. We will not encounter this parameterisation in this course, but you should be able to recognise this when someone chooses to use this way.]

$X \sim \text{Gamma}(\alpha, \theta)$ has mean α/θ, variance α/θ^2, and moment generating function $M_X(t) = (1 - t/\theta)^{-\alpha}$ for $t < \theta$.

Connection with the exponential distribution

$\text{Gamma}(1, \theta)$ simply becomes $\text{Exp}(\theta)$, the exponential distribution with rate parameter θ. This has mean $1/\theta$. On the other hand, let $E_1, \ldots, E_n$ be i.i.d. $\text{Exp}(\theta)$ random variables. Then $\sum_{i=1}^n E_i \sim \text{Gamma}(n, \theta)$, while the sample mean follows $\frac{1}{n}\sum_{i=1}^n E_i \sim \text{Gamma}(n, n\theta)$ (hint: using mgf to check).

1.2 Chi-squared distribution

Definition

If $Z_1, \ldots, Z_k$ are i.i.d. standard normal random variables. Then

$$X = \sum_{i=1}^k Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom. We write $X \sim \chi_k^2$. Note that k is a positive integer.

Connection with the Gamma distribution

As shown in Problem Set 2 Question 2, χ_k^2 is in fact Gamma$(k/2, 1/2)$, which has mean k and variance $2k$. Moreover, when $k = 2$, this becomes Gamma$(1, 1/2)$, or equivalently Exp$(1/2)$.

1.3 t-distribution

Definition* and basic properties

The t-distribution with $k > 0$ degrees of freedom (denoted t_k) has density

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2} \qquad \text{for } -\infty < t < \infty.$$

Note that here the degree of freedom k can be any positive real number. The density of a t-distribution is symmetric and bell-shaped, like that of a normal distribution. However, t-distribution has much heavier tails (see Problem set 3 Question 5(c)).

Connection with other distributions

You are not expected to memorise the above probability density function. However, you need to know how you can obtain a t_k random variable from a standard normal and an independent χ_k^2 random variable (see Problem set 3 Question 5(b)). More specifically, Let $V \sim \chi_k^2$ and $Z \sim N(0, 1)$, with Z and V being independent. Then

$$T = \frac{Z}{\sqrt{V/k}} \sim t_k.$$

t_1 distribution is also known as the Cauchy distribution, the mean of which does not exist, despite the density being symmetric.

1.4 F-distribution

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of

$$\frac{X_1/d_1}{X_2/d_2},$$

where $X_1 \sim \chi_{d_1}^2$, $X_2 \sim \chi_{d_2}^2$ and X_1 and X_2 are independent (Problem Set 4 Question 3). We denote the distribution as F_{d_1,d_2}.

If $T \sim t_k$, then $T^2 \sim F_{1,k}$. Can you show this?

1.5 *Beta distribution

Definition

The probability density function of the beta distribution, with shape parameters $\alpha, \beta > 0$ is

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \qquad \text{for } 0 \geq x \geq 1.$$

We denote this distribution as Beta(α, β).

Connection with order statistics

$\alpha = \beta = 1$: Uniform[0, 1];
$\alpha = n$, $\beta = 1$: Maximum of n independent Uniform[0, 1] random variables;
$\alpha = 1$, $\beta = n$: Minimum of n independent Uniform[0, 1] random variables;
More generally, The kth order statistic $U_{(k)}$ of a sample of size n from Uniform[0, 1] follows a Beta distribution: $U_{(k)} \sim \text{Beta}(k, n+1-k)$.

Connection with Gamma distribution

Let $X \sim \text{Gamma}(\alpha, \theta)$ and $Y \sim \text{Gamma}(\beta, \theta)$, with X and Y independent, then

$$\frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta).$$

Immediate corollary: let $X \sim \chi^2_{k_1}$ and $Y \sim \chi^2_{k_2}$, with X and Y independent. Then $X/(X+Y) \sim \text{Beta}(k_1/2, k_2/2)$.

2 Sample mean, sample variance, normal population

Let $Y_1, \ldots, Y_n$ be a random sample of size n from a population, which has mean μ and variance σ^2.

$$\text{Sample mean: } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\text{Sample variance: } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

We can show (Problem set 2 Question 1) that the sample mean $\bar{Y}$ is uncorrelated with $Y_i - \bar{Y}$ for $i = 1, \ldots, n$. Note that this holds with no distributional assumption on the population. However, $\bar{Y}$ can still, in general, be correlated with the sample variance S^2, unless the third central moment of the population is zero [$\text{Cov}(\bar{Y}, S^2) = \mu_3/n$].

Normal population

Assume now that we are sampling from a normal population. Then we have the following (wk2 lec1):

- $\bar{Y} \sim N(\mu, \sigma^2/n)$;
- $(n-1)S^2 \sim \sigma^2 \chi^2_{n-1}$;
- $\bar{Y}$ and S^2 are independent.

Question: What is the distribution of the following quantity?

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{S^2}}$$

Hint: look at Section 1.3.

3 Order statistics

Let $Y_1, \ldots, Y_n$ be a random sample, with CDF $F_{Y_1}(y)$ and PDF $f_{Y_1}(y)$. The order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are also random variables, defined by sorting the values of $Y_1, \ldots, Y_n$ in increasing order. Let's start with the first order statistic (or smallest order statistic or simply minimum of the sample): $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$.

$$\begin{aligned}
\text{CDF:} \quad F_{Y_{(1)}}(y) &= \mathbb{P}(\min\{Y_1, \ldots, Y_n\} \le y) = 1 - \mathbb{P}(\min\{Y_1, \ldots, Y_n\} > y) = 1 - \prod_{i=1}^{n} \mathbb{P}(Y_i > y), \\
&= 1 - [\mathbb{P}(Y_i > y)]^n = 1 - (1 - F_{Y_1}(y))^n, \\
\text{PDF:} \quad f_{Y_{(1)}}(y) &= \frac{d}{dy} F_{Y_{(1)}}(y) = n f_{Y_1}(y)(1 - F_{Y_1}(y))^{n-1}.
\end{aligned}$$

Next the nth order statistic (or largest order statistic or maximum): $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$.

$$\begin{aligned}
\text{CDF:} \quad F_{Y_{(n)}}(y) &= \mathbb{P}(\max\{Y_1, \ldots, Y_n\} \le y) = \prod_{i=1}^{n} \mathbb{P}(Y_i \le y) = F_{Y_1}(y)^n, \\
\text{PDF:} \quad f_{Y_{(n)}}(y) &= \frac{d}{dy} F_{Y_{(n)}}(y) = n f_{Y_1}(y) F_{Y_1}(y)^{n-1}.
\end{aligned}$$

In general, for the kth order statistic ($1 \le k \le n$), we have

$$\begin{aligned}
\text{CDF:} \quad F_{Y_{(k)}}(y) &= \mathbb{P}(Y_{(k)} \le y) = \mathbb{P}(\text{at least } k \text{ sample} \le y) = \sum_{j=k}^{n} \binom{n}{j} F_{Y_1}(y)^j (1 - F_{Y_1}(y))^{n-j}, \\
\text{PDF:} \quad f_{Y_{(k)}}(y) &= \frac{d}{dy} F_{Y_{(k)}}(y) = \lim_{h \downarrow 0} \frac{F_{Y_{(k)}}(y+h) - F_{Y_{(k)}}(y)}{h} = \lim_{h \downarrow 0} \frac{\mathbb{P}(y < Y_{(k)} \le y+h)}{h} \\
&= \lim_{h \downarrow 0} \frac{\mathbb{P}(k-1 \text{ sample} \le y, \text{ one sample between } y \text{ and } y+h, \ n-k \text{ sample} > y+h)}{h} \\
&= \lim_{h \downarrow 0} \frac{n!}{(k-1)!(n-k)!} \frac{F_{Y_1}(y)^{k-1} \big(h f_{Y_1}(y)\big) \big(1 - F_{Y_1}(y+h)\big)^{n-k}}{h} \\
&= \frac{n!}{(k-1)!(n-k)!} F_{Y_1}(y)^{k-1} f_{Y_1}(y) \big(1 - F_{Y_1}(y)\big)^{n-k}.
\end{aligned}$$

Finally, for $1 \le i < j \le n$, the joint density of $Y_{(i)}$ and $Y_{(j)}$ is

$$\begin{aligned}
f_{Y_{(i)}, Y_{(j)}}(y_1, y_2) &= \lim_{h_1, h_2 \downarrow 0} \frac{\mathbb{P}\big(y_1 < Y_{(i)} \le y_1 + h_1, y_2 < Y_{(j)} \le y_2 + h_2\big)}{h_1 h_2} \\
&= \lim_{h_1, h_2 \downarrow 0} \frac{F_{Y_1}(y_1)^{i-1} \big(h_1 f_{Y_1}(y_1)\big) \big(F_{Y_1}(y_2) - F_{Y_1}(y_1 + h_1)\big)^{j-i-1} \big(h_2 f_{Y_1}(y_2)\big) \big(1 - F_{Y_1}(y_2 + h_2)\big)^{n-j}}{h_1 h_2} \\
&\quad \times \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\
&= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f_{Y_1}(y_1) f_{Y_1}(y_2) F_{Y_1}(y_1)^{i-1} \big(F_{Y_1}(y_2) - F_{Y_1}(y_1)\big)^{j-i-1} \big(1 - F_{Y_1}(y_2)\big)^{n-j}.
\end{aligned}$$

Try convince yourself why the second equality above holds.