6. Sampling Distributions

6.1. Basic Definitions.

Definition 6.1. A set of random variables $X_1, \ldots, X_n$ are called a **random sample** of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called **independent and identical distributed** (**iid**) random variables with pdf or pmf $f(x)$.

The iid sampling model describes the (most frequently used) scenario that repeated observations are generated from the same probability distribution. Furthermore, these observations are taken in such a way that the value of one observation does not affect other observations. In other words, $X_1, \ldots, X_n$ are independent. As a result, the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

Definition 6.2. Let $X_1, \ldots, X_n$ be a random sample of size n from a population $f(x)$, and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution** of Y. In contrast, a **parameter** can be generally thought as a function of the population distribution $f(x)$.

Although simple, it is important to fully understand the above definitions and know the difference between a parameter and a statistic.

Example 6.3. Let $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$. Examples of statistics are $Y_1 = \sum_i X_i/n$, $Y_2 = \sum_i (X_i - Y_1)^2/(n-1)$. Examples of parameters are $p_1 = \int x f(x)dx = \mu$, $p_2 = \int x^2 f(x)dx = \sigma^2 + \mu^2$ and $p_3 = \int (x-\mu)^2 f(x)dx = \sigma^2$. The quantity $\sum_i (X_i - \mu)/n$ is not a statistic since it involves the unknown parameter μ!

Definition 6.4. The **sample mean** is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The **sample variance** is the statistic defined by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The **sample standard deviation** is the statistic defined by $S = \sqrt{S^2}$.

Theorem 6.5. *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then (1) $\mathbb{E}(\bar{X}) = \mu$; (2) $\mathbb{V}(\bar{X}) = \sigma^2/n$; (3) $\mathbb{E}(S^2) = \sigma^2$.*

Proof. Omitted. □

The sampling distribution of a statistic is of great importance. Often we are given a random sample and the goal is to estimate some interesting parameters of the population (which are generally unknown). We will need the theoretical form of the sample distribution in order to make inference about the parameters of the underlying population that we are interested in. In this course we focus on sampling distributions related to the normal distribution (and the reason will become clear later). However the above definition and theorem about sample mean and variance are applicable to any population distribution.

6.2. **Sampling Distributions Related to Normal.**

Theorem 6.6. *Let iid samples* $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ *, then*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathsf{N}(\mu, \sigma^2/n).$$

The result can be easily proved by noticing that $\bar{X}$ is a linear combination of independent normal random variables (see the mgf section). It follows that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$.

Definition 6.7. A random variable X is said to have a **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if the density function of X is

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & 0 \le y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$.

The Gamma distribution is very useful and has many special cases. For example, when $\alpha = 1$ it degenerates to the exponential distribution (verify). Here we are interested in another special case: the chi-square distribution.

Definition 6.8. A random variable X is said to have a **chi-square distribution** with $n \in \mathbb{N}$ **degrees of freedom** if it is a gamma-distributed random variable with parameters $\alpha = n/2$ and $\beta = 2$. We use the notation $X \sim \chi^2_n$ and we have

$$f(x) = \begin{cases} \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, & 0 \le y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Example 6.9. Let $X \sim \chi^2_n$, calculate its mgf $M_X(t)$.
By definition we have

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) \\ &= \int e^{tx}\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}dx \\ &= 2^{-n/2}((\frac{1}{2}-t)^{-1})^{n/2}\left\{\int \frac{x^{n/2-1}}{((\frac{1}{2}-t)^{-1})^{n/2}\Gamma(\frac{n}{2})}\exp\left(-\frac{x}{(\frac{1}{2}-t)^{-1}}\right)dx\right\} \\ &= (1-2t)^{-n/2} \end{aligned}$$

where the last equation comes from the fact that the quantity inside the curly brackets is just the integral over a Gamma random variable and thus equals to 1.

Theorem 6.10. *If* $Z_1, \ldots, Z_n$ *are iid* $\mathsf{N}(0,1)$ *then* $Y = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$.

Proof. We use mgf to prove the result. Let $Y_i = Z_i^2$ and its mgf is

$$\begin{aligned} M_{Y_i}(t) &= \mathbb{E}(e^{tY_i}) = \mathbb{E}(e^{tX_i^2}) = \int_{-\infty}^{\infty} e^{tx^2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{x^2}{(1-2t)^{-1}}\right)dx \\ &= (1-2t)^{-1/2}. \end{aligned}$$

So this is the mgf of χ^2_1. So we have the mgf of Y is

$$M_Y(t) = \prod_{i=1}^n M_{Y_i}(t) = (1-2t)^{-n/2}.$$

As a result, $Y = \sum_{i=1}^n X_i^2 \sim \chi^2_n$. $\square$

The reason why we introduce the chi-square distribution is due to the following fact.

Theorem 6.11. *Let $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ and S^2 the sample variance. Then*

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Furthermore, $\bar{X}$ and S^2 are independent random variables.

Proof. (following Exercise 13.83 in [1]) First, assume $X_1, \ldots, X_n \sim \mathsf{N}(0,1)$. Define an $n \times n$ matrix $\mathbf{A}^1$ by

$$A = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{2\cdot 3}} & \frac{1}{\sqrt{2\cdot 3}} & \frac{-2}{\sqrt{2\cdot 3}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & & \cdots & & \frac{1}{\sqrt{(n-1)n}} & \frac{-(n-1)}{\sqrt{(n-1)n}} \end{bmatrix}.$$

Notice that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, e.g., $\mathbf{A}$ is an orthogonal matrix. Let $\vec{Y} = \mathbf{A}\vec{X}$ and we have $\sum_{i=1}^n Y_i^2 = \vec{Y}^T\vec{Y} = \vec{X}^T\mathbf{A}^T\mathbf{A}\vec{X} = \vec{X}^T\vec{X} = \sum_{i=1}^n X_i^2$. Also note that $Y_1 = \sqrt{n}\bar{X}$.

The joint pdf of $\vec{Y}$ is

$$f_{\vec{Y}}(\vec{y}) = |\mathbf{A}^{-1}|(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}[\mathbf{A}^{-1}\vec{y}]_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right).$$

So we have $Y_1, \ldots, Y_n$ are independent. We also have $\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2$. As a result, $\sum_{i=1}^n (X_i - \bar{X})$ and $\sqrt{n}\bar{X}$ are independent. Since $\bar{X}$ and S^2 are functions of independent random variables, we have $\bar{X}$ and S^2 are independent too. Furthermore, $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$.

If $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$ we can re-define $Z_i = (X_i - \mu)/\sigma \sim \mathsf{N}(0,1)$ and it follows since $\sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2$. □

Theorem 6.6 can be used for inference of unknown μ if we know σ^2. If we do not know σ^2, then we can replace it with S^2 and consider $\sqrt{n}(\bar{X} - \mu)/S$ instead. It turns out that this quantity follows a t-distribution with $(n-1)$ degrees of freedom.

Definition 6.12. Let $Z \sim \mathsf{N}(0,1)$ and $W \sim \chi^2_\nu$. Then if Z and W are independent, the random variable $T = \frac{Z}{\sqrt{W/\nu}}$ is said to have a t-**distribution** with ν degrees of freedom. We denote $T \sim t_\nu$.

If $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, we know that $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathsf{N}(0,1)$ and $W = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Furthermore, we know that Z and W are independent. Then by the above definition we have

$$T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}.$$

So given a sample we can do inference on μ even with unknown σ^2.

Now suppose we want to compare the variances of two normal populations based on two independent random samples, one from each population. In particular, suppose we get n_1 samples $X_1, \ldots, X_{n_1}$ from population one with variance σ_1^2 and get n_2 samples $Y_1, \ldots, Y_{n_2}$ from population two with variance σ_2^2. Let S_1^2 and S_2^2 be the sample variances of the two samples respectively, then we show $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ has an F distribution based on the following definition:

Definition 6.13. Let $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ be independent random variables. Then $F = \frac{W_1/\nu_1}{W_2/\nu_2}$ is said to have an F **distribution** with (ν_1, ν_2) degrees of freedom. We denote $F \sim F_{\nu_1,\nu_2}$.

[1] Actually the matrix $\mathbf{A}$ can be any $n \times n$ orthogonal matrix with its first row equals $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$.

So we have

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{(n_1-1)S_1^2/((n_1-1)\sigma_1^2)}{(n_2-1)S_2^2/((n_2-1)\sigma_2^2)} = \frac{W_1/(n_1-1)}{W_2/(n_2-1)} \sim F_{n_1-1,n_2-1}.$$

This result will enable us to do inference about variances later.