1 Sampling distributions

In this chapter, we briefly review some concepts from MATH 61.2 regarding the probability distributions of functions of random variables (now called statistics), using techniques like moment-generating functions and the cdf technique. We prove the Lindeberg-Lévy form of the celebrated central limit theorem, as well as derive certain special distributions of sample means, sample variances, and order statistics. Finally, we introduce three new probability distributions, the χ^2 distribution, the student t distrubution, and the $\mathcal{F}$ distribution.

1.1 Distribution of the mean and the CLT

Given n random variables, $X_1, X_2, \ldots, X_n$, a **statistic** is a function $g(X_1, X_2, \ldots, X_n)$ of these random variables. The probability distribution of a statistic is called its sampling distribution.

Example. Let $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{Be}(p)$. Define T_n to be their sum, $T_n = \sum_{i=1}^n X_i$. Then, we know that $T_n \sim \text{Bin}(n, p)$. ■

Example. Let $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$. Find the distribution of $X_1 + X_2$.

Note that the support for $X_1 + X_2$ is the unit square in $\mathbb{R}^2$. The function $x_1 + x_2 = t$ corresponds to a line moving diagonally across the square from left to right. We use the cdf technique:

$$F_{X_1+X_2}(t) = \mathbf{Pr}[X_1 + X_2 \le t] = \iint_R 1 \cdot \mathrm{d}x_1 \mathrm{d}x_2, \; R = \{(x_1, x_2) \in [0, 1] : x_1 + x_2 \le t\}.$$

Notice that we can split the integral into two cases: if (1) $0 < t < 1$, then R is bounded by the two axes and $x_1 + x_2 = t$; whereas if (2) $1 \le t < 2$, then R is the unit square minus the upper right triangle bounded by $x_2 = 1, x_1 = 1$, and $x_1 + x_2 = t$. This gives us:

→ when $0 < t < 1$:

$$F_{X_1+X_2}(t) = \int_0^t \int_0^{t-x_2} 1 \cdot \mathrm{d}x_1 \mathrm{d}x_2 = \int_0^t (t - x_2)\mathrm{d}x_2 = \left(tx_2 - \frac{x_2^2}{2}\right)\Bigg|_0^t = t^2 - \frac{t^2}{2} = \frac{t^2}{2}.$$

→ when $1 \le t < 2$:

$$\begin{aligned} F_{X_1+X_2}(t) &= 1 - \int_{t-1}^1 \int_{t-x_2}^1 1 \cdot \mathrm{d}x_1 \mathrm{d}x_2 = 1 - \int_{t-1}^1 (1 - t + x_2)\mathrm{d}x_2 \\ &= 1 - \left(x_2(1-t) + \frac{x_2^2}{2}\right)\Bigg|_{t-1}^1 = \frac{1}{2} + (t-1) - \frac{(t-1)^2}{2}. \end{aligned}$$

This gives us the cdf of $X_1 + X_2$. To find its pdf, we simply differentiate and obtain

$$p_{X_1+X_2}(t) = \begin{cases} t, & \text{when } 0 < t < 1 \\ 2 - t, & \text{when } 1 \le t < 2 \end{cases}$$

as our desired pdf. ■

Consider $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Using moment-generating functions, we can determine the distribution of the so-called **sampling mean** $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, a common statistic. First recall that $\mathcal{M}_{\alpha X}(t) = \mathcal{M}_X(\alpha t)$, and $\mathcal{M}_{\sum X_i}(t) =$

$\prod_i \mathcal{M}_{X_i}(t)$ for a random variable X. Now:

$$\begin{aligned}\mathcal{M}_{\overline{X}_n}(t) = \mathcal{M}_{\frac{1}{n}\sum X_i}(t) = \mathcal{M}_{\sum X_i}\left(\tfrac{t}{n}\right) &= \prod_{i=1}^{n} \mathcal{M}_{X_i}\left(\tfrac{t}{n}\right) \\ &= \left\{\exp\left[\mu\left(\frac{t}{n}\right) + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right]\right\}^n = \exp\left[\mu t + \frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2\right], \qquad \text{substituting } \frac{t}{n} \text{ into } \mathcal{M}_{X_i}(t)\end{aligned}$$

which is the mgf of a normal random variable with mean μ and variance $\frac{\sigma^2}{n}$, so $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. In general, recall also from MATH 61.2 that if $(X_i)_{i=1}^n$ are independent normal random variables with means μ_i and variances σ_i^2, then the distribution of a linear combination of them would obey a normal distribution with a similarly linear combination of their means and variances. Symbolically,

$$(X_i)_{i=1}^n \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \implies \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

This particular quality of the sum of normal random variables being normally distributed is the **infinite divisibility** of the normal distribution. Other distributions that obey this law are the gamma distribution and the Poisson distribution.

Example. Compute the probability that the sample mean of size 10 taken from a normal population with mean 1 and variance 2 has a value between 1.2 and 3.1.

We know that the sample mean $\overline{X}_{10} \sim \mathcal{N}(1, \frac{2}{10})$. Thus,

$$\begin{aligned}\mathbf{Pr}[1.2 < \overline{X}_{10} < 3.1] &= \texttt{pnorm(3.1, mean = 1, sd = sqrt(0.2))} \\ &\quad - \texttt{pnorm(1.2, mean = 1, sd = sqrt(0.2))} \approx 0.3273,\end{aligned}$$

using R commands. ■

In practice, though, it might be that we don't know the underlying distribution of a sample $(X_i)_{i=1}^n$. This problem arises frequently when dealing with real data, which does not, in general, follow an explicitly given probability distribution. However, we have the following result:

Theorem 1.1.1: Lindeberg-Lévy CLT

Suppose $(X_i)_{i=1}^n$ are i.i.d. random variables obeying an unknown probability distribution with mean μ and variance σ^2. Then,

$$\overline{X}_n \xrightarrow{\text{d}} \mathcal{N}(\mu, \tfrac{\sigma^2}{n}),$$

where $\xrightarrow{\text{d}}$ indicates the distribution approaching $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$.

Proof. To prove this, we will use mgf's. Let $\mathcal{M}_X(t)$ be the mgf of one of these X_i's. To show that $\mathcal{M}_{\overline{X}_n}(t)$ approaches the mgf of a normal random variable with mean μ and variance $\frac{\sigma^2}{n}$ as $n \to \infty$, it suffices to show that

$$\lim_{n\to\infty} \mathcal{M}_Z(t) = \mathrm{e}^{t^2/2}, \text{ where } Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}, \text{ the standardized form of } \overline{X}_n. \tag{1.1}$$

First, we manipulate Z to make it resemble a linear combination of the X_i's:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right) = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu).$$

Substituting this into (1.1) gives us

$$\mathcal{M}_Z(t) = \mathcal{M}_{\sum(X_i-\mu)}\left(\tfrac{t}{\sigma\sqrt{n}}\right) = \prod_{i=1}^{n} \mathcal{M}_{X_i-\mu}\left(\tfrac{t}{\sigma\sqrt{n}}\right) = \left[\mathcal{M}_{X_i-\mu}\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Since we are working with exponents to the n, it might be helpful to take the logarithm of this expression. Hence, notice that to prove (1.1), it further suffices to show

$$\lim_{n\to\infty} \ln \mathcal{M}_Z(t) = \frac{t^2}{2}.$$

As a final simplification, we can make a substitution letting $h := \frac{t}{\sigma\sqrt{n}}$, so that $n = \frac{t^2}{\sigma^2 h^2}$ and $h \to 0$ as $n \to \infty$. Then:

$$\begin{aligned}
\lim_{n\to\infty} \ln \mathcal{M}_Z(t) &= \lim_{n\to\infty} \ln \left[\mathcal{M}_{X_i-\mu}\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right]^n = \lim_{n\to\infty} n \cdot \ln\left(\mathcal{M}_{X_i-\mu}\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right) \\
&= \lim_{h\to 0} \frac{t^2}{\sigma^2 h^2} \cdot \ln\left(\mathcal{M}_{X_i-\mu}(h)\right) = \frac{t^2}{\sigma^2} \lim_{h\to 0} \frac{\ln\left(\mathcal{M}_{X_i-\mu}(h)\right)}{h^2} && \text{form } \frac{0}{0}, \text{ since } \mathcal{M}_X(0) = 1 \text{ for all r.v.'s } X \\
&= \frac{t^2}{\sigma^2} \lim_{h\to 0} \frac{\frac{1}{\mathcal{M}_{X_i-\mu}(h)} \cdot \mathcal{M}'_{X_i-\mu}(h)}{2h} = \frac{t^2}{\sigma^2} \lim_{h\to 0} \frac{\mathcal{M}'_{X_i-\mu}(h)}{2h \cdot \mathcal{M}_{X_i-\mu}(h)} && \text{form } \frac{0}{0}, \text{ since } \mathcal{M}'_{X_i-\mu}(0) = \mathbf{E}[X_i - \mu] = 0 \\
&= \frac{t^2}{\sigma^2} \lim_{h\to 0} \frac{\overbrace{\mathcal{M}''_{X_i-\mu}(h)}^{\mathbf{E}[(X_i-\mu)^2] \,=\, \mathbf{var}[X_i] \,=\, \sigma^2}}{\underbrace{2 \cdot \mathcal{M}_{X_i-\mu}(h)}_{2} + \underbrace{2h \cdot \mathcal{M}'_{X_i-\mu}(h)}_{0}} = \frac{t^2}{\sigma^2} \cdot \frac{\sigma^2}{2} = \frac{t^2}{2},
\end{aligned}$$

as desired. □

The central limit theorem (CLT) is often useful in deriving approximations of very large discrete properties, where finding exact values might be theoretically impossible due to combinatorial explosion. This can be seen in the following form of the CLT, which is a corollary of the Lindeberg-Lévy CLT:

Theorem 1.1.2: De Moivre-Laplace Theorem

If $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{Be}(p)$ (i.e., independent coin flips with probability p), then

$$\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p) \xrightarrow{\text{d}} \mathcal{N}\big(np, np(1-p)\big).$$

Thus, given a random variable X obeying a binomial distribution with very large n, we can approximate

$$\mathbf{Pr}[X = k] = \mathbf{Pr}[k - \tfrac{1}{2} < X < k + \tfrac{1}{2}] \approx \mathbf{Pr}[k - \tfrac{1}{2} < W < k + \tfrac{1}{2}],$$

where W is a normally distributed random variable with the appropriate parameters. This approximation is valid whenever $n > 9 \cdot \frac{p}{1-p}$ and $n > 9 \cdot \frac{1-p}{p}$.

Example. Let $(X_i)_{i=1}^{64} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$. Find $k \in \mathbb{R}$ such that $\mathbf{Pr}\left[\sqrt[64]{\prod_{i=1}^{64} X_i} \geq k\right] = 0.10$.

Taking the logarithm of the expression inside the probability function gives us

$$\mathbf{Pr}\left[\frac{1}{64} \sum_{i=1}^{64} \ln X_i \geq \ln k\right] = 0.10,$$

which we can recognize as the sampling mean of random variables $(\ln X_i)_{i=1}^{64}$. Ideally, then, we can apply the CLT to approximate the value of k that will make this statement true, since this will give us an approximation of the distribution of $\overline{\ln X}_{64}$, but first we need to find $\mathbf{E}[\ln X_i]$ and $\mathbf{var}[\ln X_i]$. We use the cdf technique. Note that since X_i takes on values between 0 and 1, $\ln X_i$ will take on values less than 0. Hence

$$F_{\ln X_i}(t) = \mathbf{Pr}[\ln X_i \le t] = \mathbf{Pr}[X_i \le \mathrm{e}^t] = \int_0^{\mathrm{e}^t} 1 \cdot \mathrm{d}x_i, t < 0,$$

and so $p_{\ln X_i}(t) = \mathrm{e}^t$, with $t < 0$. From here we can see that $\mathbf{E}[\ln X_i] = -1$ and $\mathbf{var}[\ln X_i] = 1$. Then, by the CLT, we have $\frac{1}{64}\sum_{i=1}^{64} \ln X_i \xrightarrow{\mathrm{d}} \mathcal{N}(-1, \frac{1}{64})$. Finally, we compute

$$\begin{aligned}
\mathbf{Pr}\left[\frac{1}{64}\sum_{i=1}^{64} \ln X_i \ge \ln k\right] &= 0.10 \\
1 - \mathbf{Pr}\left[\frac{1}{64}\sum_{i=1}^{64} \ln X_i \le \ln k\right] &= 0.10 \\
\ln k &= \texttt{qnorm(0.9, mean = -1, sd = 1/8)} \\
k &\approx \mathrm{e}^{-0.84},
\end{aligned}$$

as desired. ■

1.2 Distribution of the sample variance

After the mean, perhaps the next most common statistic in considertaion is the **sampling variance**, defined as

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Before we derive its distribution, however, it is necessary to introduce another distribution—a variant of one that we have already seen before—and prove some basic results about it.

A random variable X obeys a χ^2 distribution with k degrees of freedom if it has the following pdf:

$$f(x) = \frac{1}{\Gamma(\frac{k}{2}) \cdot 2^{k/2}} \cdot x^{k/2-1}\mathrm{e}^{-x/2}, \text{ where } x, k \in (0, \infty).$$

Note that χ_k^2 is equivalent to $\mathrm{Gamma}(\frac{k}{2}, 2)$. Thus, we have the following properties, for $X \sim \chi_k^2$:

1 $\mathbf{E}[X] = \frac{k}{2} \cdot 2 = k$.

2 $\mathbf{var}[X] = \frac{k}{2} \cdot 2^2 = 2k$.

3 $\mathcal{M}_X(t) = (1-2t)^{-k/2}, t < \frac{1}{2}$.

In practice, k is usually some positive integer, but it could theoretically take on any value.

Theorem 1.2.1

If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi_1^2$. Similarly, if $(Z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then $\sum Z_i^2 \sim \chi_n^2$.

Proof. We will use mgf's. Recall that $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \sqrt{2\pi\sigma^2}$ when f is of the form $\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, i.e., a gaussian kernel.

$$\begin{aligned}\mathcal{M}_{Z^2}(t) &= \mathbf{E}\left[\mathrm{e}^{tZ^2}\right] = \int_{-\infty}^{\infty} \mathrm{e}^{tz^2} \cdot \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\mathrm{d}z = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2} + tz^2\right)\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(1-2t)z^2\right]\mathrm{d}z = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \underbrace{\exp\left[-\frac{1}{2}\left(\frac{(z-0)^2}{(1-2t)^{-1}}\right)\right]}_{\text{gaussian kernel of } \mathcal{N}(0,\frac{1}{1-2t})} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{2\pi}{1-2t}} = (1-2t)^{-1/2} = \mathcal{M}_{\chi_1^2}(t), t < \frac{1}{2}.\end{aligned}$$

The proof of the second statement follows by simply multiplying n of these mgf's together. □

A corollary of this is that $\chi_n^2 \xrightarrow{\mathrm{d}} \mathcal{N}(n, 2n)$ by an application of the CLT. We can now prove our main result:

Theorem 1.2.2: Distribution of the sample variance

Let $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$