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. Note: some examples are taken from last year's 62.1, so don't be alarmed by an example you don't recognize.

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:

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

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) $1 \leq t < 2$, 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 $1 \leq t < 2$:

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

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

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

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,

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

using R commands. ■

Nerd Interjection! Here's a brief anatomy of an R command:

`{p|d|q}{distribution}(value, params)`

→ `{p|d|q}`: Using `p` will return the cdf of the distribution evaluated at `value`, i.e., $\mathbb{P}\mathbf{r}[X <$ `value`$]$. Using `d` will return the result of evaluating the pdf instead (useful for discrete distributions). Finally, using `q` will evaluate the inverse cdf, i.e., the solution to the equation $\mathbb{P}\mathbf{r}[X < x] =$ `value`.

→ `distribution`: Pretty self-explanatory. Names are may be shortened if long (e.g., `norm` for normal and `binom` for binomial).

→ `params`: In general, the parameters you need to define the distribution will appear in the usual order that they do and you don't really need the whole `mean =` (e.g., `pnorm(0.5, 0, 1)` is standard normal and returns 0).

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:

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

as desired. □

The central limit theorem (CLT) is often useful in deriving approximations of very large discrete properties, where finding exact values might be practically 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 $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{Be}(p)$ (i.e., independent coin flips with probability p), then

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

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

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

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 $X \sim \mathrm{Bin}(10, 0.5)$. Find $\mathbb{P}\mathbf{r}[X=3]$ by approximating it first, then by finding the exact value.

Since $n = 10 > 9$, we can use the normal approximation. Thus,

$$\begin{aligned} \mathbb{P}\mathbf{r}[X=3] &= \mathbb{P}\mathbf{r}[2.5 < X < 3.5] \approx \mathbb{P}\mathbf{r}[2.5 < W < 3.5] && \text{where } W \sim \mathcal{N}(5, 2.5) \\ &= \texttt{pnorm(3.5, 5, sqrt(2.5))} - \texttt{pnorm(2.5, 5, sqrt(2.5))} \approx 0.114467. \end{aligned}$$

By the usual pdf, we simply have:

$$\mathbb{P}\mathbf{r}[X=3] = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \texttt{dbinom(3, 10, 0.5)} \approx 0.1171875,$$

as desired. ■

Perhaps a more extreme example of a combinatorial approximation is given by the following theorem:

Theorem 1.1.3: Stirling's approximation

For large values of n,

$$n! \sim \sqrt{2\pi n} \cdot n^n \cdot \mathrm{e}^{-n}.$$

That is, the limit of their ratio approaches 1 as n approaches infinity.

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

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

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

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 $\mathbb{E}[\ln X_i]$ and $\mathbb{V}\mathbf{ar}[\ln X_i]$. 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

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

and so $p_{\ln X_i}(t) = \mathrm{e}^t$, with $t < 0$. From here we can see that $\mathbb{E}[\ln X_i] = -1$ and $\mathbb{V}\mathbf{ar}[\ln X_i] = 1$. 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

$$\begin{aligned}
\mathbb{P}\mathbf{r}\left[\frac{1}{64}\sum_{i=1}^{64} \ln X_i \geq \ln k\right] &= 0.10 \\
1 - \mathbb{P}\mathbf{r}\left[\frac{1}{64}\sum_{i=1}^{64} \ln X_i \leq \ln k\right] &= 0.10 \\
\ln k &= \mathtt{qnorm(0.9, mean = -1, sd = 1/8)} \\
k &\approx \mathrm{e}^{-0.84},
\end{aligned}$$

as desired. ∎

Example. Let X be the lifetime, in years, of an LED TV. Suppose that $X \sim \mathrm{Exp}(5)$. Find the approximate value of $\mathbb{P}\mathbf{r}[\overline{X}_{64} \geq 5.4]$, and then find the exact value of this probability.

Since $(X_i)_{i=1}^{64} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(64)$, the sample mean $\overline{X}_{64} \xrightarrow{\mathrm{d}} \mathcal{N}(5, \frac{25}{64})$ by the CLT. Then,

$$\begin{aligned}
\mathbb{P}\mathbf{r}[\overline{X}_{64} \geq 5.4] &= \mathbb{P}\mathbf{r}\left[\frac{\overline{X}_{64} - 5}{5/8} \geq \frac{5.4 - 5}{5/8}\right] \approx \mathbb{P}\mathbf{r}[Z \geq 0.64] \qquad \text{where } Z \sim \mathcal{N}(0,1) \\
&= 1 - \mathtt{pnorm(0.64)} \approx 0.2610863,
\end{aligned}$$

for our approximation. To get the exact value, recall that the sum of n exponential random variables with parameter 5 is itself a random variable $W \sim \mathrm{Gamma}(n, 5)$. Thus,

$$\begin{aligned}
\mathbb{P}\mathbf{r}[\overline{X}_{64} \geq 5.4] &= \mathbb{P}\mathbf{r}\left[\frac{1}{64}\sum_{i=1}^{64} X_i \geq 5.4\right] = \mathbb{P}\mathbf{r}\left[\sum_{i=1}^{64} X_i \geq 345.6\right] \\
&= 1 - \mathtt{pgamma(345.6, 64, 5)} \approx 0.2529934
\end{aligned}$$

is our exact value. ∎

1.2 Distribution of the sample variance

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

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

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:

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

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

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

2 $\mathbb{V}\mathrm{ar}[X] = \frac{k}{2} \cdot 2^2 = 2k$.

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.

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

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

Nerd Interjection! What the hell is a "Gaussian kernel"?! Well, we know that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x = 1$$

since it's a valid pdf, so multiplying both sides by $\sqrt{2\pi\sigma^2}$ gives us

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x = \sqrt{2\pi\sigma^2},$$

and this holds for any value of μ and σ^2. Keep this sort of thing in mind, as it shows up a lot and we'll be using this technique with the gamma distribution as well later.

A corollary of this is that $\chi_n^2 \xrightarrow{\text{d}} \mathcal{N}(n, 2n)$ 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,

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

Proof. We begin by expanding the definition of S_n^2 to turn it into a more workable form.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 &= \sum_{i=1}^{n}(X_i-\mu+\mu-\overline{X}_n)^2 = \sum_{i=1}^{n}\left[(X_i-\mu)-(\overline{X}_n-\mu)\right]^2 \\
&= \sum_{i=1}^{n}\left[(X_i-\mu)^2-2(X_i-\mu)(\overline{X}_n-\mu)+(\overline{X}_n-\mu)^2\right] \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2\sum_{i=1}^{n}(X_i-\mu)(\overline{X}_n-\mu)+\sum_{i=1}^{n}(\overline{X}_n-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2(\overline{X}_n-\mu)\left(\sum_{i=1}^{n}X_i-n\mu\right)+n(\overline{X}_n-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2n(\overline{X}_n-\mu)(\overline{X}_n-\mu)+n(\overline{X}_n-\mu)^2 = \sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}_n-\mu)^2.
\end{aligned}$$

Put more neatly, this gives us

$$\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 = \sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}_n-\mu)^2. \tag{1.2}$$

But $\frac{X_i-\mu}{\sigma}\sim\mathcal{N}(0,1)$, so we can apply thm. 1 of this section to obtain $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2$. Finally, dividing both sides of equation 1.2 by σ^2,

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 &= \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-n\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2 \\
\frac{n-1}{\sigma^2}\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2}_{S_n^2} &= \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2 \\
\frac{n-1}{\sigma^2}S_n^2 &= \underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}_{\sim\chi_n^2}-\underbrace{\left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2}_{\sim\chi_1^2} && \text{since } \frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1) \\
\mathcal{M}_{\frac{n-1}{\sigma^2}S_n^2}(t) &= \frac{\mathcal{M}_{\chi_n^2}(t)}{\mathcal{M}_{\chi_1^2}(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-\frac{n-1}{2}}, && \text{taking mgf's}
\end{aligned}$$

which is the mgf of χ_{n-1}^2, so $\frac{n-1}{\sigma^2}S_n^2\sim\chi_{n-1}^2$. □

Note that for the above trick with the mgf's to work as intended, we need $\overline{X}_n$ and S_n^2 to be independent. They *are* independent, but this will be proven in MATH 62.2.

Example. Let $(X_i)_{i=1}^{10}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\mu,4)$. Find the value of $\mathbb{P}\mathbf{r}[3.6<S^2<4.2]$.

We first scale the values to fit our theorem. We then have

$$\mathbb{P}\mathbf{r}\left[\frac{9}{4}\cdot 3.6<\frac{10-1}{4}S^2<\frac{9}{4}\cdot 4.2\right] = \mathbb{P}\mathbf{r}[8.1<\chi_9^2<9.45] = \texttt{chisq}(9.45,9)-\texttt{chisq}(8.1,9)\approx 0.1272876,$$

as desired. Suppose we wanted to find $a,b\in\mathbb{R}$ such that $\mathbb{P}\mathbf{r}[a<S^2<b]=0.95$. Then, similarly, we scale the values to obtain $\mathbb{P}\mathbf{r}[\frac{9}{4}a<\chi_9^2<\frac{9}{4}b]$. We can state this equivalently as $\mathbb{P}\mathbf{r}[\chi_9^2<\frac{9}{4}a]=0.025$ and $\mathbb{P}\mathbf{r}[\chi_9^2<\frac{9}{4}b]=0.975$, which gives us

$$\frac{9}{4} \cdot a = \texttt{chisq}(0.025, 9) \qquad \frac{9}{4} \cdot b = \texttt{chisq}(0.975, 9)$$
$$a = 1.200 \qquad b = 8.4545,$$

as desired. ■

1.3 The student t and the $\mathcal{F}$ distributions

Warning for this section: in my opinion, LaTeX is not the best at rendering gnarly integrals, especially for really tiny fonts, so try to zoom in.

Now that we know how to find the mean and variance of a random sample, we can look into perhaps *standardizing* it. Recall that the standardized form of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by $\frac{X-\mu}{\sigma}$. Equivalently, we can 'standardize' the mean of a normal random sample with the following expression:

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}.$$

Note that the $\sqrt{n}$ comes from the fact that $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. We will see that the following holds:

Theorem 1.3.1: Distribution of the standardized sample mean

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

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1},$$

denoting a student t distribution with $n-1$ degrees of freedom.

Before we prove this theorem, let's make sense of this mysterious t. Suppose we have two independent random variables, $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_k$. We want the pdf of

$$T := \frac{Z}{\sqrt{W/k}}.$$

We can solve for this using the Jacobian method.

Let $t := \frac{z}{\sqrt{t/k}}$ and $v := w$. We first need to find the joint pdf of Z and W. Since they're independent, this is simply

$$f(z, w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot \frac{1}{\Gamma(\frac{k}{2}) 2^{k/2}} w^{\frac{k}{2}-1} e^{-\frac{w}{2}},$$

with (z, w) taking on values in $\mathbb{R} \times \mathbb{R}^+$. Now we need to solve for z and w in terms of t and v. Obviously, $w = v$, and we also have

$$t = \frac{z}{\sqrt{w/k}} \implies z = t \cdot \sqrt{\frac{w}{k}} = t \cdot \sqrt{\frac{v}{k}}.$$

To get the range of possible values of t and v, we note that z can be any real number, and v must always be positive, so t must likewise be any real number. The support then remains unchanged. Finally, we find the Jacobian:

$$\mathcal{J} = \det\left(\begin{bmatrix} \frac{\partial z}{\partial t} & \frac{\partial w}{\partial t} \\ \frac{\partial z}{\partial v} & \frac{\partial w}{\partial v} \end{bmatrix}\right) = \frac{\partial z}{\partial t}\frac{\partial w}{\partial v} - \frac{\partial w}{\partial t}\frac{\partial z}{\partial v} = \sqrt{\frac{v}{k}} \cdot 1 - \frac{t}{\sqrt{k}} \cdot \frac{1}{2} v^{-\frac{1}{2}} \cdot 0 = \sqrt{\frac{v}{k}},$$

which is always positive. This gives us our joint pdf of T and V:

$$\begin{aligned} g(t,v) = f\left(t\sqrt{\tfrac{v}{k}}, v\right) \cdot |\mathcal{J}| &= \frac{1}{\sqrt{\pi} \cdot \Gamma(\frac{k}{2}) \cdot 2^{\frac{k+1}{2}}} \cdot v^{\frac{k}{2}-1} e^{-\frac{v}{2}} e^{-\frac{t^2 v}{2k}} \cdot \sqrt{\frac{v}{k}} \\ &= \frac{1}{\sqrt{k\pi} \cdot \Gamma(\frac{k}{2}) \cdot 2^{\frac{k+1}{2}}} \cdot v^{\frac{k}{2}+\frac{1}{2}-1} e^{-\frac{1}{2}\left(1+\frac{t^2}{k}\right)v}. \end{aligned}$$

You might notice we've arranged the variables on the right side a little bit suggestively, and the reward comes when we integrate out v to get the pdf of t alone.

$$\begin{aligned} g(t) = \int_{-\infty}^{\infty} g(t,v)\mathrm{d}v &= \frac{1}{\sqrt{k\pi} \cdot \Gamma(\frac{k}{2}) \cdot 2^{\frac{k+1}{2}}} \underbrace{\int_{-\infty}^{\infty} v^{\frac{k}{2}+\frac{1}{2}-1} e^{-\frac{1}{2}\left(1+\frac{t^2}{k}\right)v} \mathrm{d}v}_{\text{kernel of Gamma}\left(\frac{k+1}{2}, \frac{2}{1+t^2/k}\right)} \\ &= \frac{1}{\sqrt{k\pi} \cdot \Gamma(\frac{k}{2}) \cdot 2^{\frac{k+1}{2}}} \cdot \Gamma(\tfrac{k+1}{2}) \left(\frac{2}{1+\frac{t^2}{k}}\right)^{\frac{k+1}{2}} = \frac{1}{\sqrt{k\pi}} \cdot \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \cdot \left(\frac{1}{1+\frac{t^2}{k}}\right)^{\frac{k+1}{2}}, \end{aligned}$$

with $\operatorname{supp}(T) = \mathbb{R}$.

Nerd Interjection! In case you missed it, here's the gamma kernel trick:

$$\int_0^{\infty} \frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} \mathrm{d}x = 1,$$

so multiplying both sides by $\Gamma(\alpha) \cdot \beta^{\alpha}$ gives us

$$\int_0^{\infty} x^{\alpha-1} e^{-\frac{x}{\beta}} \mathrm{d}x = \Gamma(\alpha) \cdot \beta^{\alpha}$$

no matter what the values of α and β are.

We can now introduce the **student** t distribution. A random variable X obeys a t distribution with k degrees of freedom if it has the following pdf:

$$f(x) = \frac{1}{\sqrt{k\pi}} \cdot \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \cdot \left(\frac{1}{1+\frac{x^2}{k}}\right)^{\frac{k+1}{2}}, \ x \in \mathbb{R}.$$

We write $X \sim t_k$. The t distribution is symmetric about 0, but with fatter tails than the standard normal distribution, and it has the following properties:

1. $\mathbb{E}[X] = \mathbb{E}[Z] \cdot \mathbb{E}[\sqrt{\frac{k}{W}}] = 0$, where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_k$.
2. $\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] = \mathbb{E}[T^2] \cdot \mathbb{E}[\frac{k}{W}] = \frac{k}{\Gamma(\frac{k}{2}) \cdot 2^{\frac{k}{2}}} \underbrace{\int_0^{\infty} w^{\frac{k}{2}-1-1} e^{-\frac{w}{2}} \cdot \mathrm{d}w}_{\text{kernel of Gamma}(\frac{k}{2}-1,\,2)} = \frac{k}{\underbrace{\Gamma(\frac{k}{2})}_{=(\frac{k}{2}-1)\Gamma(\frac{k}{2}-1)} \cdot 2^{\frac{k}{2}}} \cdot \Gamma(\tfrac{k}{2}-1) \cdot 2^{\frac{k}{2}-1} = \frac{k}{k-2}.$

3 The limit of the t_n pdf as n goes to infinity approaches the pdf of a standard normal random variable. That is, if $f_n(t)$ is the pdf of a student t random variable with n degrees of freedom,

$$\lim_{n\to\infty} f_n(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

We can now prove theorem 1.3.1:

Proof. Notice that since $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Then,

$$\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}} = \underbrace{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}_{\sim\mathcal{N}(0,1)} \Bigg/ \underbrace{\sqrt{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}}_{\sim\sqrt{\frac{\chi^2_{n-1}}{n-1}}},$$

so this has a t_{n-1} distribution, where $Z = \frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$, $W = \frac{(n-1)S_n^2}{\sigma^2}$, and $k = n-1$. □

Example. Let $\overline{X}_n$ and S^2 be the sample mean and variance of a normal sample of size 16 with mean μ and variance σ^2. Find the probability that μ is within 2 standard deviations of $\overline{X}_n$.

The probability that we want is $\mathbb{P}\mathbf{r}[\overline{X}_n - 2\sqrt{S^2} < \mu < \overline{X}_n + 2\sqrt{S^2}]$. With some manipulation, this becomes:

$$\begin{aligned}\mathbb{P}\mathbf{r}[\overline{X}_n - 2\sqrt{S^2} < \mu < \overline{X}_n + 2\sqrt{S^2}] &= \mathbb{P}\mathbf{r}[-2\sqrt{S^2} < \overline{X}_n - \mu < 2\sqrt{S^2}] = \mathbb{P}\mathbf{r}[-2S < \overline{X}_n - \mu < 2S] \\ &= \mathbb{P}\mathbf{r}\left[-2 < \frac{\overline{X}_n-\mu}{S} < 2\right] = \mathbb{P}\mathbf{r}\left[-8 < \underbrace{\frac{\overline{X}_n-\mu}{S/4}}_{\sim t_{15}} < 8\right] \\ &= \texttt{pt(8, 15)} - \texttt{pt(-8, 15)} \approx 0.9999991,\end{aligned}$$

as desired. ■

We introduce one more new distribution to close off this section. We say that a continuous random variable obeys an $\mathcal{F}$ distribution if it is defined as the ratio of two χ^2 random variables. More formally, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ be independent. Then,

$$\frac{X_1/k_1}{X_2/k_2} \sim \mathcal{F}_{k_1,k_2}.$$

To find the pdf of this distribution, we use the Jacobian method again.

Given these two χ^2 random variables, let $u = \frac{x_1/k_1}{x_2/k_2}$ and $v = x_2$. By independence, the joint pdf of X_1 and X_2 is simply

$$f(x_1,x_2) = \frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}} \cdot e^{-\frac{x_1+x_1}{2}} \cdot x_1^{\frac{k_1}{2}-1} \cdot x_2^{\frac{k_2}{2}-1},$$

with (x_1,x_2) taking on values in $\mathbb{R}^+ \times \mathbb{R}^+$. Now we solve for x_1 and x_2 in terms of u and v. Obviously $x_2 = v$, and

$$x_1 = k_1 u \cdot \frac{x_2}{k_2} = k_1 u \cdot \frac{v}{k_2} = \frac{k_1}{k_2} uv.$$

We can see that the support remains unchanged. Solving for the Jacobian:

$$\mathcal{J} = \det\left(\begin{bmatrix} \frac{\partial x_1}{\partial u} & \frac{\partial x_2}{\partial u} \\ \frac{\partial x_1}{\partial v} & \frac{\partial x_2}{\partial v}\end{bmatrix}\right) = \frac{\partial x_1}{\partial u}\frac{\partial x_2}{\partial v} - \frac{\partial x_1}{\partial u}\frac{\partial x_2}{\partial v} = \frac{k_1}{k_2} v \cdot 1 - \frac{k_1}{k_2} u \cdot 0 = \frac{k_1}{k_2} v.$$

Thus, we can solve for the joint pdf of u and v:

$$\begin{aligned} g(u,v) &= \frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}}\left(\frac{k_1}{k_2}uv\right)^{\frac{k_1}{2}-1}\cdot v^{\frac{k_2}{2}-1}\cdot\exp\left[-\frac{1}{2}\left(\frac{k_1}{k_2}uv+v\right)\right]\cdot\frac{k_1}{k_2}v \\ &= \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}\cdot\frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}}\cdot u^{\frac{k_1}{2}-1}\cdot v^{\frac{k_1+k_2}{2}-1}\exp\left[-\frac{v}{2}\left(1+\frac{k_1}{k_2}u\right)\right]. \end{aligned}$$

Finally, we integrate out v (gamma kernel incoming!)

$$\begin{aligned} g(u) &= \int_0^\infty\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}\cdot\frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}}\cdot u^{\frac{k_1}{2}-1}\cdot v^{\frac{k_1+k_2}{2}-1}\exp\left[-\frac{v}{2}\left(1+\frac{k_1}{k_2}u\right)\right]\mathrm{d}v \\ &= \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}\cdot\frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}}\cdot u^{\frac{k_1}{2}-1}\int_0^\infty\underbrace{v^{\frac{k_1+k_2}{2}-1}\exp\left[-\frac{v}{2}\left(1+\frac{k_1}{k_2}u\right)\right]}_{\text{kernel of Gamma}(\frac{k_1+k_2}{2},\frac{2}{1+u\frac{k_1}{k_2}})}\mathrm{d}v \\ &= \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}\cdot\frac{1}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})\cdot 2^{\frac{k_1+k_2}{2}}}\cdot u^{\frac{k_1}{2}-1}\cdot\Gamma(\tfrac{k_1+k_2}{2})\cdot\left(\frac{2}{1+u\frac{k_1}{k_2}}\right)^{\frac{k_1+k_2}{2}} \\ &= \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}\cdot\frac{\Gamma(\frac{k_1+k_2}{2})}{\Gamma(\frac{k_1}{2})\cdot\Gamma(\frac{k_2}{2})}\cdot u^{\frac{k_1}{2}-1}\cdot\left(\frac{1}{1+u\frac{k_1}{k_2}}\right)^{\frac{k_1+k_2}{2}}, v>0, \end{aligned}$$

which is the pdf of an $\mathcal{F}$ random variable.

1.4 Order statistics

With these new distributions under our belts, we now briefly discuss a statistic that you may have encountered before in MATH 61.2. Recall that the **order statistics** of a random sample $(X_i)_{i=1}^n$ are the values arranged in ascending order, denoted by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. We refer to $X_{(i)}$ as the i^{th} order statistic, and $X_{(1)}$ and $X_{(n)}$ as the minimum and maximum respectively. Note that each order statistic is not in general independent from the rest.

Theorem 1.4.1: Distribution of the i^{th} order statistic

Let $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$ having pdf $f(x)$ and cdf $F(x)$ for all X_i. Then, the pdf of $X_{(i)}$ is given by

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}\cdot[1-F(x)]^{n-i}\cdot f(x),\ 1\le 1\le n.$$

In particular, $f_{X_{\min}}(x) = n\cdot[1-F(x)]^{n-i}\cdot f(x)$ and $f_{X_{\max}}(x) = n\cdot[F(x)]^{i-1}\cdot f(x)$.

We will not prove this theorem in all its generality, but we give the following informal argument:

Suppose that the i^{th} order statistic $X_{(i)}$ has value x, and that all order statistics are independent. This corresponds to the pdf of X_i being evaluated at x, giving us the $f(x)$ term. Then, there are $i-1$ order statistics *below* $X_{(i)}$, all with value $< x$, which correspond to the $[F(x)]^{i-1}$ term. Similarly, there are $n-i$ order statistics *above* $X_{(i)}$, all with value $> x$, which correspond to the $[1-F(x)]^{n-i}$ term. Finally, we divide out the possible permutations of the $i-1$ terms and the $n-i$ terms, giving us the $\frac{n!}{(i-1)!(n-i)!}$ term. By independence, we simply multiply all these terms together, giving us the distribution.

Sometimes finding the pdf of the order statistics is just a simple "plug-and-chug", but other techniques may also be used, as seen in the following example:

Example. Let $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \operatorname{Exp}(2)$. Find the pdf of $\hat{\theta} = X_{\min}$.

Since the X_i obey an exponential distribution, $f_X(x) = \frac{1}{2}\exp(-\frac{x}{2})$, with $x > 0$. Then, we can use the cdf technique:

$$\begin{aligned} F_{\hat{\theta}}(t) &= \mathbb{P}\mathbf{r}[\hat{\theta} \le t] = 1 - \mathbb{P}\mathbf{r}[\hat{\theta} > t] = 1 - \mathbb{P}\mathbf{r}\left[\bigcap_{i=1}^n X_i > t\right] \\ &= 1 - \left[\int_t^\infty \frac{1}{2}\mathrm{e}^{-\frac{y}{2}}\,\mathrm{d}x\right]^n = 1 - \left[\lim_{a\to\infty} \mathrm{e}^{-\frac{x}{2}}\Big|_t^a\right]^n = 1 - \mathrm{e}^{-\frac{nt}{2}}, \end{aligned}$$

so the pdf of $\hat{\theta}$ is $f(x) = \frac{n}{2}\exp(-\frac{nx}{2})$, $x > 0$. ■

2 Estimators

In this chapter, we look at problems of estimating parameters of certain unknown probability distributions, problems which arise naturally given samples from experimental data. We begin with two methods, maximum likelihood estimation (MLE) and the method of moments (MOM). We then look at a few properties of parameters; specifically, we formalize what it means for an estimator to be unbiased, efficient, and consistent, and we study (but not prove!) a theorem known as the Cramer-Rao lower bound. Finally, we look at problems of interval estimation.

2.1 MLE and MOM

Probability theory is responsible for giving a simplified model of a random process. However, in most cases, the 'best' (i.e., most accurate and precise) model for capturing this process is unknown. It is our role as researchers to identify this model, hence the need for statistics.

To formalize this problem, consider the following situation. Suppose we have a random sample $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} f(x;\theta)$. It may be that f is unknown, or that θ is unknown (or both!), but we restrict ourselves only to the case where f comes from a known family of distributions, and wee need to approximate θ. Here's a simple example:

Example. Say we want to measure an object, with the true measurement being some (unknown) θ. What we can do is make n independent measurements of the object, giving us a sample $(X_i)_{i=1}^n$. To get the 'best' estimate for θ, we want an estimate so that the difference between the true value and all our measurements is minimized, so we can define the least squares metric: $L(\theta) \coloneqq \sum_i (X_i - \theta)^2$. Since we want the value of θ that minimizes this function, let's take its partial derivative with respect to θ and set it to be 0, then solve for θ:

$$\frac{\partial L}{\partial \theta} = 2\sum_{i=1}^n (X_i - \theta) \cdot (-1) = -2\sum_{i=1}^n (X_i - \theta).$$

This gives us $\theta = \sum_i X_i / n = \overline{X}_n$. ■

We now define a **point estimator** of a parameter θ to be a statistic $T(X_1, \ldots, X_n)$. We denote point estimators as $\hat{\Theta}$. If a realized sample is given, i.e., $X_i = x_i$ for all i, then $T(x_1, \ldots, x_n)$ is a real number called a **point estimate** of θ. We denote it as $\hat{\theta}$. We consider two main methods of estimation.