Econ 240A (1st Half)
Section 2: Fall 2018
Friday, September 7

Fengshi Niu[*]

Contents

[*] Thanks to Mikkel Sølvsten and Matias D. Cattaneo for letting me use their old notes as an inspiration.

1 Recap of expectation

Remember the *the expectation of a random variable* X, denoted $\mathbb{E}X$, is given by

$$\begin{aligned}\mathbb{E}(X) &= \int XdP = \int X(\omega)dP(\omega) = \int_{\mathbb{R}} xdF_X(x) \\ &= \begin{cases} \sum_x x \cdot P\left(X = x\right) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x \cdot f_X\left(x\right) \cdot dx & \text{if } X \text{ is continuous} \end{cases},\end{aligned}$$

provided that the sum or integral exists. If $\mathbb{E}|X| = \infty$, then we say that $\mathbb{E}X$ does not exists. Observe that in the definition we are using *Lebesgue-Stieltjes integral.* This integral is in fact a generalization of the Riemann Integral and circumvents the pdf, which is an awkward concept for mixed distributions (distributions with discrete and continuous components). Mixed random variables have cdfs that are differentiable everywhere except at a countable set of points. This generalization of integration allows for a complete treatment. Intuitively, this integral just put together both definitions of expectation (when discrete and when continuous) by applying the corresponding definition to each portion of the average (i.e., the integral).

Let's now turn to the example where the underlying random variable is not discrete nor continuous, but mixed.

Example 1.1. *(*Expectation of a Mixed Random Variable*) Let* $Z \sim \mathcal{N}(0,1)$ *and consider the (mixed) random variable* X *defined as*

$$X = Z \cdot \left(1 - \mathbb{I}\left\{-k \leq Z \leq k\right\}\right),$$

where $k > 0$*. Then the cdf and pdf are respectively given by (why?),*

$$F_X\left(x\right) = \begin{cases} \Phi\left(x\right) & \textit{if } x < -k \\ \Phi\left(-k\right) & \textit{if } -k \leq x < 0 \\ \Phi\left(k\right) & \textit{if } 0 \leq x < k \\ \Phi\left(x\right) & \textit{if } k \leq x \end{cases}.$$

Then to compute the expected value we note

$$\begin{aligned}\mathbb{E}\left[g\left(X\right)\right] &= \int_{-\infty}^{\infty} g\left(x\right) \cdot dF_X\left(x\right) \\ &= \int_{-\infty}^{-k} g\left(x\right) \cdot \phi\left(x\right) \cdot dx + g\left(0\right) \cdot \left(\Phi\left(k\right) - \Phi\left(-k\right)\right) + \int_{k}^{\infty} g\left(x\right) \cdot \phi\left(x\right) \cdot dx.\end{aligned}$$

In the previous example we illustrated how to compute expectations for mixed random variables. In the next exercise we compute the expected value for another non-invertible transformation of a random variable, which leads to a mixed random variable. This exercise is hard but it is worth to try especially if you want to see another example of a distribution that is not continuous nor discrete.

Exercise 1.1. *(*Left Censoring*) Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and let $Y = \max\{X, k\}$. Compute the expected value of Y. (Hint: derive the cdf of Y, then compute the pdf of y and finally calculate the expected value.) It is important to note that the random variable Y is not continuous nor discrete and hence the computation of its expected value involves both integration and summation.*

As it can be seen, the previous example and exercise presented two cases of a latent model, which is often used in econometrics.

Before we move on to the next section to discuss a special class of expectations, which is generically known as moments, we present the following exercise that gives an additional interpretation of expected value.

Exercise 1.2. *(*Minimizing Distance*) Let X be a random variable with $\mathbb{E}\left[|X|\right] < \infty$. Show that:*

$$\mathbb{E}\left[X\right] = \arg\min_a \mathbb{E}\left[(X-a)^2\right].$$

The previous exercise says that, intuitively, if we try to select a value a such that minimizes $\mathbb{E}\left[(X-a)^2\right]$, which can be interpreted as an average loss from trying to predict a realization of X, the result is in fact $\mathbb{E}\left[X\right]$. This, in turn, suggest that $\mathbb{E}\left[X\right]$ can be interpreted as the best "guess" under this quadratic loss function and when we are allowed to pick a constant. Moreover, observe that in this case, $\mathbb{E}\left[(X-a^*)^2\right] = \mathbb{E}\left[(X-E\left[X\right])^2\right] = \mathbb{V}ar\left[X\right]$.

Notice that the previous exercise suggests a way of constructing an estimator which, of course, depends on the loss function chosen. For example, if we use absolute deviation as the loss we will obtain a different answer. More precisely,

$$\begin{aligned}\mathbb{E}\left[X\right] &= \arg\min_a \mathbb{E}\left[(X-a)^2\right], \\ \text{Median}(X) &= \arg\min_a \mathbb{E}\left|X-a\right|.\end{aligned}$$

2 Moments and moment generating functions

From the definition of expected value, we have the following two classes of moments:

1. If we let $g\left(X\right) = X^n$, then we obtain the n-th **moment** of X which is denoted by

$$\mu'_n = \mathbb{E}\left[X^n\right].$$

2. If we let $g\left(X\right) = \left(X - \mu'_1\right)^n$, then we obtain the n-th **central moment** of X which is denoted by

$$\mu_n = \mathbb{E}\left[\left(X-\mu'_1\right)^n\right] = \mathbb{E}\left[(X-\mu)^n\right].$$

The second central moment of a random variable X is known as variance (of X) and it gives a measure of spread of a distribution around its mean.

Definition 2.1. *The* ***variance*** *of a random variable is defined as*

$$\mathbb{V}ar\left[X\right] \equiv \mu_2 = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[(X-\mathbb{E}\left[X\right])^2\right].$$

Because the measurement units on the variance is the square of the original unit, sometimes we will be dealing with the positive squared-root of the variance, which is known as standard deviation. Two important properties of the variance are stated in the next Theorem.

Theorem 2.1. *Let X be a random variable with finite variance, then we have:*

1. $\mathbb{V}ar\left[X\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2$.
2. *For any $a, b \in \mathbb{R}$, $\mathbb{V}ar\left[a+bX\right] = b^2 \cdot \mathbb{V}ar\left[X\right]$.*

Proof. To see (1), observe that

$$\begin{aligned}
\mathbb{V}ar\left[X\right] &= \mathbb{E}\left[(X-\mu)^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mu + \mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[2X\mu\right] + \mathbb{E}\left[\mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2.
\end{aligned}$$

To see (2), recall that $\mathbb{E}\left[a+bX\right] = a + \mathbb{E}\left[bX\right]$ and thus

$$\begin{aligned}
\mathbb{V}ar\left[a+bX\right] &= \mathbb{E}\left[((a+bX) - \mathbb{E}\left[a+bX\right])^2\right] \\
&= \mathbb{E}\left[(a+bX-a-b\cdot\mathbb{E}\left[X\right])^2\right] \\
&= b^2 \cdot \mathbb{E}\left[(X-\mathbb{E}\left[X\right])^2\right] \\
&= b^2 \cdot \mathbb{V}ar\left[X\right],
\end{aligned}$$

which concludes the proof. □

At this point it is important to mention other functions of moments that may appear later in this or other classes. This is done in the next definition.

Definition 2.2. *Let X be a random variable. The* ***standard deviation*** *of X is given by*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathbb{V}ar\left[X\right]}.$$

The ***skewness*** *of X is given by*

$$\alpha_3 = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{(\mathbb{V}ar\left[X\right])^{\frac{3}{2}}} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}}.$$

*Finally, the **kurtosis** of X is given by*

$$\alpha_4 = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{(\mathbb{V}ar\,[X])^2} = \frac{\mu_4}{(\mu_2)^2}.$$

Now we present a very handy function that, when exists, can be used to derive very useful properties of random variables.

Definition 2.3. *Let X be a random variable. The **moment generating function (mgf)** of X (or F_X), is the function $M_X : \mathbb{R} \to \bar{\mathbb{R}}_+$ given by*

$$M_X(t) = \mathbb{E}\left[\exp\{tX\}\right].$$

Observe that although this function is well-defined, we have to discuss two cases. When the mgf is finite in some neighborhood of zero (that is, there exists $h > 0$ such that for all $t \in (-h, h)$, $M_X(t) < \infty$), we will say it exists; while if $M_X(t) = \infty$ we will say that it does not exists. This is unfortunate distinctions highlight that the usefulness of this function is somehow limited to the cases when it exits. Although in this class, we will deal only with such cases, for completeness we introduce in the next subsection a different generating function that actually always exists and enjoy very similar properties to those the mgf has, giving more general and powerful results.

In general, generating functions have at least three uses: first, it can be used to compute moments explicitly in a systematic way; second, it can be used to obtain the distribution of functions of random variables; and finally, it can be used to obtain limiting distributions of functions of random variables. In this class we will discuss the first two uses, which are presented in the next theorems.

Theorem 2.2. *If the random variable X has mgf $M_X(t) < \infty$ in a neighborhood of zero, then*

$$\mathbb{E}\left[X^n\right] = \left.\frac{d^n}{dt^n} M_X(t)\right|_{t=0}.$$

Proof. We prove this result for the case of continuous random variable (as always, the discrete case is analogous provided that the integral sign is replaced by the summation sign). Assuming we can differentiate under the integral sign, we have

$$\begin{aligned}
\frac{d^n}{dt^n} M_X(t) &= \left.\frac{d^n}{dt^n} \int_{-\infty}^{+\infty} \exp\{tx\} \cdot f_X(x) \cdot dx\right|_{t=0} \\
&= \int_{-\infty}^{+\infty} \left(\frac{d^n}{dt^n} \exp\{tx\}|_{t=0}\right) \cdot f_X(x) \cdot dx \\
&= \int_{-\infty}^{+\infty} \left(x^n \exp\{tx\}|_{t=0}\right) \cdot f_X(x) \cdot dx \\
&= \int_{-\infty}^{+\infty} x^n \cdot f_X(x) \cdot dx \\
&= \mathbb{E}\left[X^n\right],
\end{aligned}$$

and the result follows. □

The next theorem is particularly useful since it gives a way of characterize distributions of (transformations of) random variables by using the mgf.

Theorem 2.3. *Let X and Y be two random variables with cdfs $F_X(x)$ and $F_Y(y)$, respectively. Assuming all moments exists for these two random variables, we have:*

1. *If X and Y have bounded support, then*

   $$X \overset{d}{=} Y \iff F_X(z) = F_Y(z), \text{ for all } z \in \mathbb{R} \iff \mathbb{E}[X^n] = \mathbb{E}[Y^n] \text{ for all } n \in \mathbb{N}_0 = \{0, 1, 2, ...\}$$

2. *If the moment generating function exists and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then*

   $$F_X(z) = F_Y(z), \text{ for all } z \in \mathbb{R} \left(\iff X \overset{d}{=} Y\right)$$

At this point it is important to work out a complete example using the results reported in this section. The following exercise ask you to work through these definitions. Problem Set 2 will also require some similar work for the case of the Gamma distribution.

Exercise 2.1. *Suppose that $X \sim Poisson(\lambda)$.*

1. *Verify that $\mathbb{E}[X] = \lambda = \mathbb{V}ar[X]$.*
2. *Show that $M_X(t) = \exp\{\lambda(e^t - 1)\}$.*
3. *Assuming $\{X_i\}_{i=1}^n \sim Poisson(\lambda_i)$ independently, show that $Y_n \equiv \sum_{i=1}^n X_i \sim Poisson(\sum_{i=1}^n \lambda_i)$.*

Characteristic Function

As we discussed before, the moment generating function is very useful whenever it exists. As some may have noticed, the moment generating function is the two-sided Laplace transform of the density of the random variable and its main problem is that it integrates an unbounded function and hence it is natural to expect that the integral may be infinite in many cases. An alternative generating function uses a different transform:

Definition 2.4. *The **characteristic function (ch.f.)** of X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ given by*

$$\varphi_X(t) = \mathbb{E}[\exp\{itX\}] = \mathbb{E}[\cos\{tX\}] + i \cdot \mathbb{E}[\sin\{tX\}].$$

As some of you may have recognized, the ch.f. is the Fourier transform of the density and since it is clearly bounded, this function is always finite. Thus, although this function involves using some basics of complex analysis, it additional power makes it extremely useful for proving asymptotics results as well as many other things. In this class, however, we will not use this function and we will only rely on the mgf.

In case you want to practice a bit, here is an easy exercise:

Exercise 2.2. *(*MGF AND CH.F. OF NORMAL DISTRIBUTION*) Let $X \sim \mathcal{N}\left[\mu,\sigma^2\right]$. Show that*

$$M_X\left(t\right) = \exp\left\{t\mu + \frac{1}{2}\sigma^2 t^2\right\}, \quad \textit{and} \quad \varphi_X\left(t\right) = \exp\left\{it\mu - \frac{1}{2}\sigma^2 t^2\right\}.$$

This result will be of great use when deriving asymptotic distributions of estimators.

3 Families of distributions

At the end of Casella and Berger (2002), pp. 621-627, it can be found a very good summary of the most common distributions used in Statistics. In particular, page 627 includes an amazing figure linking all the distributions that we will be studying in this class. We strongly suggest that you take a look to this table and make sure that you can recover the argument for the most relevant families (Normal, Gamma, etc.). In this Section we will present and discuss the exponential family, which is one of the most used families in both theoretical and empirical Statistics.

At this point it instructive to review our main goal in this class. Recalling our discussion in the first section: our goal is to learn as much as possible from the data and in order to do this, we need to impose some assumptions. The most common assumption is that the observed data are realizations of some random variable with an unknown distribution function. Then, our goal is to learn as much as possible about this distribution. In mathematical terms, we have

$$X \sim P_{\theta_0} \in \mathcal{P} = \left\{P_\theta : \theta \in \Theta\right\}.$$

At this point, however, this assumptions are usually insufficient in order to learn about the data generating process. The main reason of this failure is that this model $\mathcal{P}$ is to "big". Therefore, one way to proceed, is to reduce the dimension of this model. For example, we can assume that

$$\begin{aligned}\mathcal{P}_1 &= \left\{\text{all probabilities } P_\theta \text{ with finite first moment} : \theta \in \Theta\right\},\\ \mathcal{P}_2 &= \left\{\text{all normal distributions, } P_\theta : \theta \in \Theta\right\}.\end{aligned}$$

The first model, $\mathcal{P}_1$, is nonparametric, while the second model, $\mathcal{P}_2$, is parametric. Clearly, $\mathcal{P}_2 \subset \mathcal{P}_1$, and of course $\mathcal{P}_2$ is much smaller. In many applications, our goal is to be as general as possible while restricting ourself to the parametric world; that is, when the model $\mathcal{P}$ is known up to a finite number of parameters.

The *exponential family* is a name for a class of models that encompasses lots of parametric families, e.g. $\mathcal{P}_2$. In particular, we have

$$f\left(x|\,\theta\right) = h\left(x\right) \cdot c\left(\theta\right) \cdot \exp\left\{\sum\nolimits_{k=1}^{K} w_k\left(\theta\right) \cdot t_k\left(x\right)\right\},$$

where $K \geq 1$ and

$$\begin{aligned} h &: \mathbb{R} \to \mathbb{R}, \\ c &: \Theta \to \mathbb{R}_{++}, \\ w_k &: \Theta \to \mathbb{R}, \text{ for all } k = 1, 2, ..., K, \\ t_k &: \mathbb{R} \to \mathbb{R}, \text{ for all } k = 1, 2, ..., K. \end{aligned}$$

This family has many interesting properties, many of which we will discuss in this class during the following weeks. In particular, the next theorem presents one of such properties. The proof of this theorem will be done later, when we discuss the properties of the log-likelihood function. For now we concentrate on the result.

Theorem 3.1. *If X is a random variable with pdf or pmf belonging to the exponential family, then*

$$\begin{aligned} \mathbb{E}\left[\sum\nolimits_{k=1}^{K} \left(\frac{\partial}{\partial \theta_j} w_k(\theta)\right) \cdot t_k(X)\right] &= -\frac{\partial}{\partial \theta_j} \log\{c(\theta)\}, \\ \mathbb{V}ar\left[\sum\nolimits_{k=1}^{K} \left(\frac{\partial}{\partial \theta_j} w_k(\theta)\right) \cdot t_k(X)\right] &= -\frac{\partial^2}{\partial \theta_j^2} \log\{c(\theta)\} - \mathbb{E}\left[\sum\nolimits_{k=1}^{K} \left(\frac{\partial^2}{\partial \theta_j^2} w_k(\theta)\right) \cdot t_k(X)\right]. \end{aligned}$$

An important definition that will be proved to be very useful later is given below.

Definition 3.1. *The* ***support*** *of a distribution is the set*

$$supp(X; \theta) = \{x \in \mathbb{R} : f_X(x | \theta) > 0\} \subset \mathbb{R}.$$

Observe that in particular for the case of the exponential family, the support is not parameter dependent since we have

$$\text{supp}(X; \theta) = \{x \in \mathbb{R} : f_X(x | \theta) > 0\} = \{x \in \mathbb{R} : h(x) > 0\} = \text{supp}(X),$$

this fact is of course a necessary but not sufficient condition that characterizes any density belonging to this family.

We close this Section by presenting an exercise where we apply these results.

Exercise 3.1. *Suppose that $X \sim Poisson(\lambda)$.*

1. *Show that X belongs to the exponential family.*
2. *Verify that $\mathbb{E}[X] = \lambda = \mathbb{V}ar[X]$ using the theorem for exponential families.*
3. *Describe the support of X.*

4 Inequalities

In this Section we review three of the most important (univariate) inequalities: Markov's, Chebyshev's and Jensen's inequalities. Later, probably in our next meeting, we will discuss other inequalities relating more than one random variable. These are also very useful inequalities, especially for proving theoretical results.

Theorem 4.1. *(*Markov's Inequality*) Let X be a random variable with density function $f_X(x)$, and let $g(\cdot)$ be a non-negative valued function. Then*

$$\mathbb{P}[g(X) \geq a] \leq \frac{\mathbb{E}[g(X)]}{a}, \textit{ for any value of } a > 0.$$

Remember you proved this inequality in a problem set 1.

Corollary 4.1. *(*Chebyshev's inequality*) Let X be a random variable with finite first moment, $\mu = \mathbb{E}[X]$ and $\sigma = \sqrt{\mathbb{E}\left[(X-\mu)^2\right]}$. Then*

$$\mathbb{P}[|X-\mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}, \textit{ for any value of } \varepsilon > 0.$$

Before we move to the last inequality, we leave one important question as exercise. This will lead us to a rudimentary notion of convergence of random variables that will be revisited later.

Exercise 4.1. *Using the Chebyshev's Inequality, what happens if $\sigma^2 \to 0$ with the random variable X?*

The last inequality that we will discuss is known as Jensen's Inequality and is a direct consequence of convexity of function and linearity of expectation.

Theorem 4.2. *(*Jensen's Inequality*) Let X be a random variable, and $g(\cdot)$ a convex function. Assume that $\mathbb{E}[|g(X)|] < \infty$. Then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Proof. By convexity of $g(x)$ we have

$$g(\mathbb{E}[X]) + g'(\mathbb{E}[X]) \cdot (x - \mathbb{E}[X]) \leq g(x)$$

and taking expectations both sides we have

$$\mathbb{E}\left[g(\mathbb{E}[X]) + g'(\mathbb{E}[X]) \cdot (X - \mathbb{E}[X])\right] \leq \mathbb{E}[g(X)],$$

and hence

$$g(\mathbb{E}[X]) + g'(\mathbb{E}[X]) \cdot \mathbb{E}[X - \mathbb{E}[X]] \leq \mathbb{E}[g(X)],$$

and thus $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$, which concludes the proof. □