HG
Oct. 14

ECON 4130 H14

Exercises for the seminar week 41

Introduction: Review of multidimensional distributions

(This summarizes what you basically need to know about joint distributions in this course.)

In the lectures I have only talked about joint two-dimensional distributions, but everything mentioned about them generalizes straightforward to higher dimensions. For the sake of completeness I review the basics here (Rice is a bit vague on this):

If $X_1, X_2, \ldots, X_n$ are rv's, their joint cdf is defined by

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1 \cap X_2 \le x_2 \cap \ldots \cap X_n \le x_n)$$

The corresponding joint pdf or pmf is defined by

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} P(X_1 = x_1 \cap \ldots \cap X_n = x_n) & \text{discrete case} \\ \dfrac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n) & \text{continuous case} \end{cases}$$

In the continuous case we have: [**Note** that everything below hold for discrete distributions as well, replacing integrals by sums and pdf's by pmf's.]

The cdf is determined by the pdf by

$$F(x_1, x_2, \ldots x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_{n-1}} \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n) du_n du_{n-1} \cdots du_2 du_1$$

(or replacing the integrals by sums in the discrete case)

calculated by starting from the innermost integral and working out step by step outwards each single integral.

The marginal joint pdf for any sub-collection of rv's is obtained by integrating away all the other variables. To simplify notation consider the four rv's, X, Y, Z, U with pdf $f(x, y, z, u)$. For example, the marginal pdf's of Y and (X, Z, U) are respectively

$$f_1(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z, u) dx dz du \quad \text{and} \quad f_2(x, z, u) = \int_{-\infty}^{\infty} f(x, y, z, u) dy$$

(and correspondingly in the discrete case replacing integrals by sums)

The expectation of any function, $g(X,Y,Z,U)$, of X,Y,Z,U can be found as before

$$E\left[g(X,Y,Z,U)\right]=\begin{cases}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y,z,u)\,f(x,y,z,u)\,dxdydzdu & \text{continuous case}\\ \sum_{\text{all } u}\sum_{\text{all } z}\sum_{\text{all } y}\sum_{\text{all } x} g(x,y,z,u)\,f(x,y,z,u) & \text{discrete case}\end{cases}$$

whenever the integral (or sum) exists.

The conditional distribution of *Y* keeping (X,Z,U) fixed to the numbers (x,z,u) is determined by the conditional pdf defined (just as in the two-dimensional case) by

(1) $$f(y\mid x,z,u)=\frac{f(x,y,z,u)}{f_2(x,z,u)}$$

Note that this determines a one-dimensional distribution of *Y* where x,z,u appear as parameters.

The conditional expected value of *Y* in (1), sometimes called the regression function of *Y* with respect to (X,Z,U), and the conditional variance of *Y* in (1) are functions (!) of (x,z,u)

$$\mu(x,z,u)=E(Y\mid x,z,u),\qquad \sigma^2(x,z,u)=\operatorname{var}(Y\mid x,z,u)$$

The law of total expectation (also called "the law of double expectation") holds in general (same proof as in the two-dimensional case):

$$E(Y)=E\left[E(Y\mid X,Z,U)\right]\overset{def}{=}E\left[\mu(X,Z,U)\right]$$

$$\operatorname{var}(Y)=E\left[\operatorname{var}(Y\mid X,Z,U)\right]+\operatorname{var}\left[E(Y\mid X,Z,U)\right]\overset{def}{=}E\left[\sigma^2(X,Z,U)\right]+\operatorname{var}\left[\mu(X,Z,U)\right]$$

If (and only if) $X_1,X_2,\ldots,X_n$ are **independent**, the joint pdf (as well as the joint cdf) can be factorized into a product of the marginal pdf's (cdf's):

(2)
$$f(x_1,x_2,\ldots,x_n)=f_{X_1}(x_1)\,f_{X_2}(x_2)\cdots f_{X_n}(x_n)\quad\text{and}\quad F(x_1,x_2,\ldots,x_n)=F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$$

Similarly, the expectation of the product $X_1X_2\cdots X_n$ factorizes under independence:

(3) $$E(X_1X_2\cdots X_n)=E(X_1)E(X_2)\cdots E(X_n)$$

If $X_1,X_2,\ldots,X_n$ are independent, the mgf (if it exists) of the sum, $S=X_1+X_2+\cdots+X_n$, can also be factorized into the product of the individual mgf's:

(4) $$M_S(t)=E\left[e^{tS}\right]=M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t)$$

which follows directly from (3) as shown in the lecture for the case $n = 2$.

A remark on modelling: To model a joint pdf (or pmf) like (e.g.) $f(x, y, z, u)$ directly is often difficult because of our common lack of intuition on the complete joint behaviour. Since it is usually easier to model one-dimensional distributions than multidimensional ones, the task is often accomplished by decomposing the joint pdf (pmf) into a product of one-dimensional pdf's (pmf's) – which is always possible due to (1). For example, in the 2-dimensional case, using (1), we have[1]

$$f(x, y) = f(y \mid x) f(x)$$

and in the 4-dimensional case, using (1) several times,

$$\begin{aligned} f(x, y, z, u) &= f(y \mid x, z, u) f(x, z, u) = f(y \mid x, z, u) f(x \mid z, u) f(z, u) = \cdots \\ &= f(y \mid x, z, u) f(x \mid z, u) f(z \mid u) f(u) \end{aligned}$$

Note that, in the special case that *Y* is considered the response (e.g., an endogenous variable), and *X, Z, U*, explanatory (e.g., exogenous) variables, it is common to stop with the first equality for modelling purposes.

A nice example of this conditioning principle you can find in the exercise (no-seminar week 40) on the ROSCA in Nairobi concerning the distribution of (V, X). There the marginal distribution of *V* is modelled as discrete uniform over $1, 2, \ldots, n$, and the conditional distribution of *X*, given $V = v$, is modelled as binomial $(v - 1, p)$. From this, if needed, we get the full joint pmf of (V, X):

$$f(v, x) = \begin{cases} \frac{1}{n} \cdot \binom{v-1}{x} p^x (1-p)^{v-1-x} & \text{for } x = 0, 1, \ldots, v-1 \text{ and } v = 2, 3, \ldots, n \\ 1 & \text{for } x = 0 \text{ and } v = 1 \\ 0 & \text{otherwise} \end{cases}$$

We would need this, e.g., to derive the marginal distribution of *X*, which is slightly complicated and not binomial(!). Luckily, in this case, we don't need the marginal distribution of *X* to derive the expected value, $E(X)$. Much simpler is to use the law of total expectation as in the exercise. So, in this case we don't need to bother about the (complicated) joint distribution at all to answer our questions of interest.

[1] In the following expressions I have used the function symbol, *f, generically* in the sense that the various *f*'s represent different functions. It is the structure of the arguments that determines which function we are talking about. This artefact is sometimes used in mathematical texts to simplify notation.

Exercise 1

Let $X_1, X_2, \ldots, X_n$ be *iid* and each exponentially distributed, $X_i \sim \exp(\lambda)$. Use (3) above to prove that for any *n*, the mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is exactly gamma-distributed. Identify the parameters in this distribution expressed by λ and n.

[**Hint**. Identify first the gamma-distribution for $S = \sum_{i=1}^{n} X_i$ using (3). Then find the distribution of $\bar{X} = \frac{1}{n} S$.]

Exercise 2

Rice chap. 4, no. 83 (cf. exercise 4 - no-seminar week 40)

Exercise 3

Let $X_1, X_2, \ldots, X_n$ be *iid* and each normally distributed, $X_i \sim N(\mu, \sigma^2)$, where $\mu = E(X_i), \quad \sigma^2 = \mathrm{var}(X_i)$. In this exercise we will compare three estimators of σ^2:

- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $\hat{\sigma}_1^2 = S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$
- $\hat{\sigma}_2^2 = S_2^3 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n+1}S^2$

a. Put $V = \sum_{i=1}^{n}(X_i - \bar{X})^2$. Using the linearity of the $\sum$ - operator, show that

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

[**Hint:** Remember that sums (like integrals) have linearity properties like, for example,

$$\sum_{i=1}^{n}(a + bx_i + cy_i) = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} y_i$$

where *a*, *b*, *c* are numbers not depending on *i*.]

b. Use supplementary exercise 4 (seminar week 38) and the fact that $\frac{V}{\sigma^2}$ is χ^2_{n-1}-distributed to show that $\hat{\sigma}^2 = S^2$ is unbiased (i.e., $E(S^2) = \sigma^2$), and has variance

$$\text{var}(\hat{\sigma}^2) = \text{var}(S^2) = \frac{2\sigma^4}{n-1}$$

c. Show that both $\hat{\sigma}_k^2, k = 1,2$, are biased downwards, i.e., $E(\hat{\sigma}_k^2) = c_k\sigma^2, \quad k = 1,2$, where both $c_k < 1$. Find c_1, c_2 and explain why both estimators satisfy, $E(\hat{\sigma}_k^2) \to \sigma^2$ when $n \to \infty$. (This property is usually expressed by saying that both estimators are *asymptotically unbiased*).

d. Gross variance (also called "mean squared error").

There are many different ways to compare estimators. In the basic course you learned to compare *unbiased* estimators by the much used criterion:

> For two different *unbiased* estimators of the same parameter, choose the one with smallest variance.

This criterion is useless in our situation where some of the estimators are biased. However, it is easy to generalize it by using (instead of variance) the *mean squared estimation error*. Let $\hat{\theta}$ be an estimator of some unknown parameter, θ. Then the mean squared estimation error (also called "gross variance" or "brutto-varians" in Norwegian) is defined by

$$\text{GVar}(\hat{\theta}) \overset{def}{=} E\left[(\hat{\theta} - \theta)^2\right]$$

(i) Explain why $\text{GVar}(\hat{\theta}) = \text{Var}(\hat{\theta})$ whenever $\hat{\theta}$ is unbiased.

(ii) Show in general that, when the variances exist, we have

(4) $$\text{GVar}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = \text{Var}(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2$$

[**Hint:** Add and subtract $E(\hat{\theta})$ inside $(\hat{\theta} - \theta)^2 = \left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right]^2$, and execute the squaring.]

Our extended criterion is now

(5) For two different estimators (not necessarily unbiased) for the same parameter, choose the one with smallest mean squared error.

e. Show, using (4), that $\text{GVar}(\hat{\sigma}_1^2) < \text{GVar}(\hat{\sigma}^2)$.

Hence $\hat{\sigma}_1^2$ is preferable to $\hat{\sigma}^2$ according to criterion (5).

[**Note** that $\hat{\sigma}_1^2$ is the same as the so called *maximum likelihood estimator*, as you will see later in the course.]

f. Defining an estimator, $\tilde{\sigma}_c^2 = cS^2$, for any constant $c > 0$, we obtain a whole class of potential estimators of σ^2. Show that *c* minimizing $\text{GVar}(\tilde{\sigma}_c^2)$ is given by

$$c = \frac{n-1}{n+1}$$

Hence, $\tilde{\sigma}_c^2 = \hat{\sigma}_2^2$ is the optimal estimator of σ^2 in this class - according to criterion (5).

[**Hint:** Minimize the expression you get with respect to *c* after using (4) on $\text{GVar}(\tilde{\sigma}_c^2)$.]

g. Prove the following result (which implies that all three estimators for σ^2 are consistent):

> Let $\hat{\theta}_n, n = 1, 2, \ldots$ be a sequence of asymptotically unbiased estimators for θ with variances that converge to zero when $n \to \infty$. Then $\hat{\theta}_n$ is a consistent estimator for θ (i.e., $\hat{\theta}_n \underset{n\to\infty}{\overset{P}{\longrightarrow}} \theta$).

[**Hint:** Show first, using Markov's inequality, that, if $\text{GVar}(\hat{\theta}_n) \underset{n\to\infty}{\to} 0$, then $\hat{\theta}_n \underset{n\to\infty}{\overset{P}{\longrightarrow}} \theta$. Then use (4) to deduce the result asked for.]

[**Note.** In econometrics we operate with several concepts of "biasedness/unbiasedness" of an estimator. In this course you learn two concepts, i.e., "(un)biasedness" and "(in)consistency": If $\hat{\theta}_n$ is an estimator of θ, we say that

- $\hat{\theta}_n$ is unbiased (biased) if $E(\hat{\theta}_n) = \theta$ $(E(\hat{\theta}_n) \neq \theta)$
- $\hat{\theta}_n$ is consistent (inconsistent) if $\underset{n\to\infty}{\text{plim}}(\hat{\theta}_n) = \theta$ $(\underset{n\to\infty}{\text{plim}}(\hat{\theta}_n) \neq \theta)$

These two concepts are not equivalent. Unbiasedness *does not* in general imply consistency, and consistency *does not* in general imply unbiasedness. Unbiasedness is a property of $\hat{\theta}_n$ for a single given *n*, while consistency is a property of the whole sequence, $\hat{\theta}_n, \ n = 1, 2, \ldots$ On

the other hand, consistency has wider applicability than unbiasedness because of the continuity property (if $\hat{\theta}_n$ is consistent for θ, then $g(\hat{\theta}_n)$ is consistent for $g(\theta)$ whenever $g(x)$ is a continuous function), a property not shared by the unbiased-concept (even if $E(\hat{\theta}_n) = \theta$, then quite often $Eg(\hat{\theta}_n) \neq g(\theta)$ - except when *g* is linear ($g(x) = a + bx$)).]