LECTURE NOTES
GRA6039 ECONOMETRICS WITH PROGRAMMING
AUTUMN 2020

EMIL A. STOLTENBERG

1. Lecture 1, August 24, 2020

In this lecture we cover some probability, introduce random variables, talk about sums and the law of large numbers. Relevant reading is Math refresher B in Wooldridge (2019) and the scanned pages from Allen (2003).

1.1. **Sets and probability.**

A *sample space* Ω is a collection of all possible *outcomes* ω of an experiment, for example $\Omega = \{H, T\}$, $\Omega = \{1, 2, 3, 4, 5, 6\}$, or $\Omega = \{\omega \colon \omega \in [0, \infty)\}$. Note the so-called set-builder notation, for example

$$\Omega = \{\omega \colon \omega \in \mathbb{N}, \omega \leq 6\} = \{1, 2, 3, 4, 5, 6\}.$$

In words: $\Omega =$ all ω such that ω is a natural number, and ω is smaller than or equal to 6. Subsets A of Ω are called *events* (well, given some technical conditions that will not bother us in this course).

Some operations on events. Let A and B be events/sets

$$\begin{aligned}
&\text{The } \textit{union} \text{ of } A \text{ and } B \text{ is the set } A \cup B = \{\omega \colon \omega \in A \,\text{or}\, \omega \in B\}.\\
&\text{The } \textit{intersection} \text{ of } A \text{ and } B \text{ is the set } A \cap B = \{\omega \colon \omega \in A \,\text{and}\, \omega \in B\}.\\
&\text{The } \textit{difference} \text{ of } A \text{ and } B \text{ is the set } A \setminus B = \{\omega \colon \omega \in A \,\text{and}\, \omega \notin B\}.\\
&\text{The } \textit{complement} \text{ of } A \text{ is the set } A^c = \{\omega \colon \omega \,\text{is not in}\, A\}.
\end{aligned} \tag{1}$$

In words (draw Venn diagrams!): The set $A \cup B$ consists of all elements ω that are in A or in B (or in both). The set $A \cap B$ consists of all elements ω that are in both A and in B. The set $A \setminus B$ consists of all elements ω that are in A and not in B, in fact $A \setminus B = A \cap B^c$. The set A^c consists of all elements ω that do not belong to A.

There is also a set called the *empty set*, denoted $\emptyset$. This is the set that has no members, we may write $\emptyset = \{\}$. Here is a fact: The empty set is a subset of all sets, that is $\emptyset \subset A$ for any set A. (To see this: Assume that $\emptyset$ is not a subset of A. Then $\emptyset$ must have a least one member that is not in A. But $\emptyset$ has no members.) Two sets A and B whose intersection is the empty set, that is $A \cap B = \emptyset$, are called *disjoint*.

Date: October 9, 2020.

Theorem 1.1. *For any three events A, B, and C defined on a sample space $\mathcal{X}$,*

$$\begin{aligned}
&\textit{Commutativity}: && A \cup B = B \cup A,\\
& && A \cap B = B \cap A;\\
&\textit{Associativity}: && A \cup (B \cup C) = (A \cup B) \cup C,\\
& && A \cap (B \cap C) = (A \cap B) \cap C;\\
&\textit{Distributive laws}: && A \cap (B \cup C) = (A \cap B) \cup (A \cap C),\\
& && A \cup (B \cap C) = (A \cup B) \cap (A \cup C);\\
&\textit{De Morgan's laws}: && (A \cup B)^c = A^c \cap B^c,\\
& && (A \cap B)^c = A^c \cup B^c.
\end{aligned}$$

Proof. Optional exercise. □

For more on sets and related stuff, you could, for example, have a look at Papineau (2012) or Hammack (2020) (these books are not part of the curriculum).

Definition 1.2. (PROBABILITY). Suppose that Ω is a sample space, and that $\mathcal{A}$ is the collection of all the events in Ω. A probability Pr is a function whose domain is $\mathcal{A}$, that obeys the following axioms:

(i) $\Pr(A) \geq 0$ for all events A;
(ii) $\Pr(\Omega) = 1$;
(iii) For all sequences $(A_n)_{n\geq 1}$ of events such that $A_n \cap A_m = \emptyset$ whenever $n \neq m$ (pairwise disjoint),
$$\Pr(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \Pr(A_n).$$

These are known as the Kolmogorov axioms. Notice that this definition tells us what rules a probability function has to obey, not what particular probability function is the correct one in a given experiment.

Here are some properties of probability functions.

Proposition 1.3. *Let* Pr *be a probability function, and A and B are events in Ω. Then*

(a) $\Pr(\emptyset) = 0$.
(b) $\Pr(A) \leq 1$.
(c) $\Pr(A) = 1 - \Pr(A^c)$.
(d) $\Pr(B \setminus A) = \Pr(B) - \Pr(A \cap B)$.
(e) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.
(f) *If $A \subset B$ then* $\Pr(A) \leq \Pr(B)$.

Proof. In class and perhaps as homework. □

Definition 1.4. (CONDITIONAL PROBABILITY). If A and B are events and $\Pr(B) > 0$, the conditional probability of A given B, written $\Pr(A \mid B)$, is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

This definition is quite intuitive (again, draw a Venn diagram): It treats B as the new sample space, and computes the fraction of B that intersects A. Suppose we have a population where 10% are smokers, 20% of the population are above 60 years old, and 5% of the population are smokers *and* over sixty. We then have $\Pr(\text{smoker}) = 1/10$, $\Pr(\text{over } 60) = 1/5$, and $\Pr(\text{smoker and over } 60) = 1/20$. A person is sampled at random from the population, and this person happens to be over sixty. What is the probability that this person is a smoker? We compute

$$\Pr(\text{smoker} \mid \text{over } 60) = \frac{\Pr(\text{smoker and over } 60)}{\Pr(\text{over } 60)} = \frac{1/20}{1/5} = \frac{1}{4}.$$

The point is that when we get to know that the person is over 60, that is, *given that* the person is over 60, we treat all people over sixty as our new population.

Proposition 1.5. *Let B be an event with $\Pr(B) > 0$. Then the function*

$$A \mapsto \Pr(\,\cdot \mid B),$$

is a probability function.

Proof. In class or optional homework. □

Definition 1.6. (Independence). Two events A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

If A and B are not independent, they are said to be dependent.

If the event A and B are independent, then A and B^c are independent, also A^c and B^c are independent. The proof of this is a nice exercise in the use of Theorem 1.1 and Proposition 1.3: Assume that A and B are independent events, then

$$\begin{aligned}
\Pr(A^c \cap B^c) &\overset{\text{De Morgan's}}{=} \Pr((A \cup B)^c) \overset{\text{Prop. 1.3(c)}}{=} 1 - \Pr(A \cup B) \\
&\overset{\text{Prop. 1.3(e)}}{=} 1 - \Pr(A) - \Pr(B) + \Pr(A \cap B) \\
&\overset{\text{Independence}}{=} 1 - \Pr(A) - \Pr(B) + \Pr(A)\Pr(B) \\
&= (1 - \Pr(A))(1 - \Pr(B)) \overset{\text{Prop. 1.3(c)}}{=} \Pr(A^c)\Pr(B^c),
\end{aligned}$$

which shows that A^c and B^c are independent.

[xx perhaps include the Law of total probability and Bayes' theorem xx]

1.2. **Random variables and distribution functions.** In many experiments it is easier to deal with, and we might be more interested in, a summary variable than with the original probability. A coin is tossed three times, there are $2^3 = 8$ possible outcomes,

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

But what we are interested in is

$$X = \#\text{ number of heads} \in \{0, 1, 2, 3\}. \tag{2}$$

Notice that X is a function from Ω to $\{0,1,2,3\}$: The space Ω is its *domain*, while $\{0,1,2,3\}$ is it *range*. We have

ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Definition 1.7. A random variable is a function from the sample Ω into the real numbers.

Suppose that the coin above is fair (equal probability for heads and tails), then

$$\Pr(HHH) = \Pr(HHT) = \cdots = \Pr(TTT) = \frac{1}{8}.$$

Note that

$$\begin{aligned}
X^{-1}(\{0\}) &= \{\omega \in \Omega \colon X(\omega) = 0\} = \{TTT\};\\
X^{-1}(\{1\}) &= \{\omega \in \Omega \colon X(\omega) = 1\} = \{TTH, THT, HTT\};\\
X^{-1}(\{2\}) &= \{\omega \in \Omega \colon X(\omega) = 2\} = \{HHT, HTH, THH\};\\
X^{-1}(\{3\}) &= \{\omega \in \Omega \colon X(\omega) = 3\} = \{HHH\}.
\end{aligned}$$

Writing $\{X = x\}$ for the more cumbersome $\{\omega \in \Omega \colon X(\omega) = x\}$ – which is standard! – we see that

$$\Pr(X=0) = \frac{1}{8}, \quad \Pr(X=1) = \frac{3}{8}, \quad \Pr(X=2) = \frac{3}{8}, \quad \Pr(X=3) = \frac{1}{8}.$$

In this sense, the random variable X *induces* a probability function, P_X say,

$$P_X(B) = \Pr X^{-1}(B) = \Pr\{\omega \in \Omega \colon X(\omega) \in B\},$$

on $\{0,1,2,3\}$ for all events B in $\{0,1,2,3\}$, for example $B = \{0\}$, or $B = \{0,1\}$, etc. That is, Pr is a probability function on Ω, while via the random variable X we get a probability function P_X in $\{0,1,2,3\}$. We say that P_X is the *distribution* of X and write

$$X \sim P_X.$$

If P_X is the normal distribution with mean μ and variance σ^2 we typically just write $X \sim \mathrm{N}(\mu, \sigma^2)$, if it is the Poisson distribution with mean λ, we write $X \sim \mathrm{Poisson}(\lambda)$, and so on.

1.3. **The summation symbol.** Let $X_1, \ldots, X_n$ be n observations, data points, random variable, numbers. Here is a definition: For integers $k \leq n$,

$$\sum_{i=k}^{n} X_i = X_k + X_{k+1} + \cdots + X_{n-1} + X_n. \tag{3}$$

For example, if $k = 1$ and $n = 4$, then $\sum_{i=1}^4 X_i = X_1 + X_2 + X_3 + X_4$. In some situations we might also write

$$\sum_{i=1}^{n} X_i = \sum_{1 \leq i \leq n} X_i = \sum_{i \in \{1,\ldots,n\}} X_i = \sum_{i \in A} X_i,$$

given that $A = \{1, \dots, n\}$. Let's say we want to sum over the numbers $1, 3, 5, 7, 9$, we can define $B = \{\text{odd numbers between 0 and 10}\} = \{1, 3, 5, 7, 9\}$, then $\sum_{j \in B} X_j = X_1 + X_3 + X_5 + X_7 + X_9$.

Again, let $X_1, \dots, X_n$ be n observations, and a and b are some constants, for example $a = 2.34$ and $b = -3.45$. Use the definition in (3),

$$\begin{aligned}\sum_{i=1}^{n}(aX_i + b) &= (aX_1 + b) + \cdots + (aX_n + b) \\ &= aX_1 + \cdots + aX_n + \underbrace{b + \cdots + b}_{n \text{ of these}} \\ &= a(X_1 + \cdots + X_n) + nb \\ &= a\sum_{i=1}^{n} X_i + nb.\end{aligned}$$

We see that constants 'go outside the sum'. By being constant we mean that they do not change with i.

1.4. **Miscellaneous.** A type of sums that appear from time to time, are the *telescoping sums*: If we have $n+1$ numbers $a_0, a_1, \dots, a_{n-1}, a_n$, then

$$\sum_{i=1}^{n}(a_i - a_{i-1}) = a_n - a_0,$$

is a telescoping sum. To see this, try a small n (always a good idea to understand sums!), say $n = 4$, then

$$\begin{aligned}\sum_{i=1}^{4}(a_i - a_{i-1}) &= (a_1 - a_0) + (a_2 - a_1) + (a_3 - a_2) + (a_4 - a_3) \\ &= \cancel{a_1} - a_0 + \cancel{a_2} - \cancel{a_1} + \cancel{a_3} - \cancel{a_2} + a_4 - \cancel{a_3} = a_4 - a_0.\end{aligned}$$

Here is a somewhat advanced example where a telescoping sum appears, and where we use many of the rules in Proposition 1.3. Suppose $A_1, A_2, \dots$ are events such that

$$A_1 \subset A_2 \subset A_3 \subset \cdots,$$

that is $(A_n)_{n \geq 1}$ is an increasing sequence of events. Let $A = \cup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup A_3 \cup \cdots$. Then probability functions are continuous in the sense that

$$\lim_{n \to \infty} \Pr(A_n) = \Pr(A). \tag{4}$$

This is not evident, and has to be proved. Define the sets

$$B_1 = A_1 \setminus A_0, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus A_2, \dots,$$

where we take $A_0 = \emptyset$. Notice that these sets are disjoint, that is $B_i \cap B_j = \emptyset$ whenever $i \neq j$. Importantly,

$$\begin{aligned}\bigcup_{n=1}^{\infty} B_n &= \bigcup_{n=1}^{\infty}(A_n \setminus A_{n-1}) = \bigcup_{n=1}^{\infty}(A_n \cap A_{n-1}^c) \\ &= (\bigcup_{n=1}^{\infty} A_n) \cap (\bigcup_{n=1}^{\infty} A_{n-1}^c) = A \cap (\bigcap_{n=1}^{\infty} A_{n-1})^c = A \cap \emptyset^c = A.\end{aligned}$$

Here we use Theorem 1.1, the Distributive laws for the third equality, and De Morgan's laws for the fourth equality. Then

$$\begin{aligned}\Pr(A) &= \Pr(\bigcup_{j=1}^{\infty} B_j) = \sum_{j=1}^{\infty} \Pr(B_j) = \lim_{n\to\infty} \sum_{j=1}^{n} \Pr(A_j \setminus A_{j-1}) \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} \{\Pr(A_j) - \Pr(A_{j-1})\} = \lim_{n\to\infty} \{\Pr(A_n) - \Pr(A_0)\} = \lim_{n\to\infty} \Pr(A_n).\end{aligned}$$

The first equality uses the result just above; the second equality is Definition 1.2(iii), using that the B_js are disjoint; the fourth equality is Proposition 1.3(d), and that $A_j \cap A_{j-1} = A_{j-1}$ because $A_{j-1} \subset A_j$; the fifth equality is what we just learned about telescoping sums; and the last equality is $A_0 = \emptyset$, and that $\Pr(\emptyset) = 0$ by Proposition 1.3(a).

If $A_1 \supset A_2 \supset A_3 \supset \cdots$ is a decreasing sequence of events, then

$$\lim_{n\to\infty} \Pr(A_n) = \Pr(\bigcap_{n=1}^{\infty} A_n). \tag{5}$$

To prove this, consider the sets $B_n = A_1 \setminus A_n$ for $n = 1, 2, \ldots$. Note that $B_n \subset B_{n+1}$ for all n. Now use (4) and Proposition 1.3(d).

Suppose X is a random variable with the uniform distribution on $[0, 1]$. That is, for any interval (a, b) in $[0, 1]$ with $a < b$,

$$\Pr(X \in (a, b)) = b - a.$$

What is the probability that $X = x$ for some $x \in [0, 1]$? We can use (5) to compute this probability. Let $A_n = \{X \in (x - 1/n, x + 1/n)\}$ for $n = 1, 2, \ldots$. We then have

$$\{X = x\} = \bigcap_{n=1}^{\infty} A_n.$$

Use this and (5) to compute the probability that $X = x$.

---

2. Lecture 2, August 31, 2020

In this lecture we'll talk about cumulative distribution functions, densities, independent random variables, expectation, and variance. Relevant reading is Math refresher B (called Appendix B in the sixth edition) in Wooldridge (2019) and the scanned pages from Allen (2003).

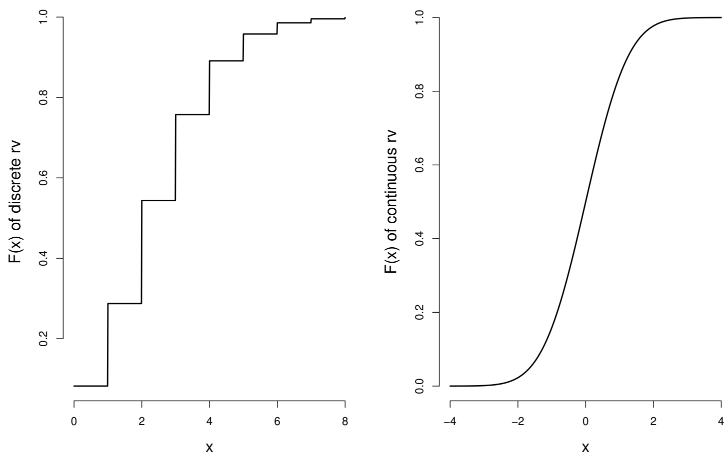

FIGURE 1. The cumulative distribution function of the Poisson distribution with mean 2.5 (left) and the of the standard normal distribution (right).

For more probability, see for example Casella and Berger (2002, ch. 1), Grimmet and Stirzaker (2001), Jacod and Protter (2012), or Shiryaev (1996) (these books are not part of the curriculum).

2.1. **Cumulative distribution functions.** The *cumulative distribution function* (cdf.) F of a random variable X is

$$F(x) = \Pr(X \leq x), \tag{6}$$

Theorem 2.1. *A function F is a cumulative distribution function if and only if it has the following properties*

(i) *$F(x)$ is nondecreasing, i.e. $F(x) \leq F(y)$ whenever $x \leq y$;*
(ii) $\lim_{x\to-\infty} F(x) = 0$, *and* $\lim_{x\to\infty} F(x) = 1$;
(iii) *$F(x)$ is right continuous, that is for each x_0, we have* $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

Proof. This theorem can be proved using the definition in (6) as well as the axioms in Definition 1.2. Not part of the curriculum. □

A *discrete* random variable is a random variable that takes its values in a set that can be listed. Examples of such sets are $\{0, 1\}$, $\{0, 1, 2, 3\}$, $\{0, 1, 2, \ldots\}$, and $\{0, 1/4, 1/2, 3/4, 1\}$. A *discrete* random variable has a cumulative distribution function with jumps, meaning that there are points x at which

$$F(x) - F(x - \delta) > 0,$$

however small you choose $\delta > 0$. Let's look at the cdf. of the random variable X from Lecture 1 (see eq. (2)) to see what this means. Recall that X takes its values in $\{0, 1, 2, 3\}$ and has distribution

$$\Pr(X = 0) = \frac{1}{8}, \quad \Pr(X = 1) = \frac{3}{8}, \quad \Pr(X = 2) = \frac{3}{8}, \quad \Pr(X = 3) = \frac{1}{8}.$$

The cdf. F of X is given by

$$F(x) = \Pr(X \leq x) = \begin{cases} 0, & -\infty < x < 0, \\ 1/8, & 0 \leq x < 1, \\ 1/2, & 1 \leq x < 2, \\ 7/8, & 2 \leq x < 3, \\ 1, & 3 \leq x < \infty. \end{cases}$$

If you make a drawing of this function, you'll see that it jumps at $x = 0$, $x = 1$, $x = 2$, and $x = 3$. For example at $x = 2$, we see that for $0 < \delta < 1$,

$$F(2) - F(2 - \delta) = \frac{1}{2} - \frac{1}{8} = \frac{3}{8} = \Pr(X = 2),$$

thus $F(x)$ makes a jump of size $\Pr(X = 2) = 3/8$ at $x = 2$.

A *continuous* random variable has a cdf. F with no such jumps, that is for each x and for any $\varepsilon > 0$, we can find a $\delta > 0$ such that

$$|F(x) - F(x - \delta)| < \varepsilon.$$

Interpretation: X is a continuous random variable if it can take any value in a subset of $\mathbb{R}$, and no single value has a positive probability of occurring. A normally distributed random variable X (with mean μ and variance σ^2) is continuous: Its cdf. is

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\big\{ -\frac{1}{2\sigma^2}(z-\mu)^2 \big\} \,\mathrm{d}z.$$

A random variable U with the uniform distribution on $[a, b]$ is continuous. Its cdf. is

$$F_U(x) = \begin{cases} 0, & -\infty < x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & b \leq x < \infty. \end{cases}$$

2.2. **Densities.** The density of a discrete random variable X is $f_X(x) = \Pr(X = x)$. We often call this the *probability mass function* (pmf.) of X, when X is discrete. If, for example $X \sim \text{Poisson}(\lambda)$, its pmf. is

$$f_X(x) = \frac{1}{x!}\lambda^x \exp(-\lambda), \quad x = 0, 1, 2, \ldots,$$

for $\lambda > 0$. For $x = 0, 1, 2, \ldots$, the cdf. is

$$F_X(x) = \Pr(X \leq x) = \sum_{z=0}^{x} f_X(z) = \sum_{z=0}^{x} \frac{1}{z!}\lambda^z \exp(-\lambda).$$

For a continuous random variable X, with a continuous cdf. $F(x)$, there is a function $f(x)$, called the *probability density function* of X, such that

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(z) \,\mathrm{d}z.$$

Using the Fundamental Theorem of Calculus (if f is continuous), we have that

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = F'(x) = f(x),$$

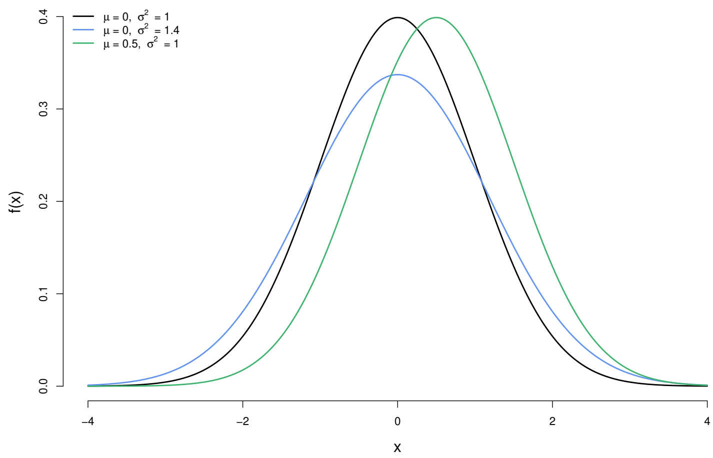

FIGURE 2. The probability density function $f(x)$ in (7) with various values for the mean μ and variance σ^2.

with $\mathrm{d}F(x)/\mathrm{d}x = F'(x)$ just being two different ways of writing the derivative with respect to x.

The pdf. of a normally distributed random variable X with mean μ and variance σ^2 is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}. \tag{7}$$

This is the famous 'bell curve' as depicted in Figure 2.

You can go from density functions to cumulative distribution functions. In fact, any function $f(x)$ such that

$$f(x) \geq 0, \text{ for all } x, \quad \text{and} \quad \sum_x f(x) = 1 \quad \text{or} \quad \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1,$$

is the pmf. or pdf. of a random variable, and $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ is its cdf. (replace the integral by a sum in the discrete case). Consider for example the function

$$f(x) = \begin{cases} \theta x^{\theta-1}, & \text{for } 0 \leq x \leq 1, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for some } \theta > 0.$$

Then

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^1 \theta x^{\theta-1}\,\mathrm{d}x = x^\theta \Big|_0^1 = 1.$$

The function $F(x)$ defined by $F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z = \int_0^x \theta z^{\theta-1}\,\mathrm{d}z$ is then the cumulative distribution function of a random variable. If we call this random variable X, then for $0 \leq a < b \leq 1$, for example

$$\Pr(a < X \leq b) = F(b) - F(a) = \int_a^b \theta x^{\theta-1}\,\mathrm{d}x = b^\theta - a^\theta,$$

is the probability that X takes its value in the interval $(a, b] \subset [0, 1]$.

2.3. **Independent random variables.** If $X_1, \ldots, X_n$ are n random variables, the joint cumulative distribution function of the vector $(X_1, \ldots, X_n)$ is

$$F(x_1, \ldots, x_n) = \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Assume that $n = 2$, that is (X_1, X_2). If $F(x_1, x_2)$ is continuous, then

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(z_1, z_2)\, \mathrm{d}z_2\, \mathrm{d}z_1,$$

where by the Fundamental Theorem of Calculus,

$$f(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F(x_1, x_2),$$

is the joint density of the random vector $(X_1, X_2)'$.

The random variables $X_1, \ldots, X_n$ are *independent* if

$$\begin{aligned} F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) &= \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= \Pr(X_1 \leq x_1) \cdots \Pr(X_n \leq x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n). \end{aligned} \tag{8}$$

for all $x_1, \ldots, x_n$. Here $F_{X_1, \ldots, X_n}(x_1, \ldots, x_n)$ is the joint cdf. of $(X_1, \ldots, X_n)$ while F_{X_i} is the cdf. of X_i, for $i = 1, \ldots, n$. The definition in (8) can also be stated in terms of densities. Suppose $X_1, \ldots, X_n$ are random variables with densities $f_{X_1}, \ldots, f_{X_n}$, then $X_1, \ldots, X_n$ are independent if

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n). \tag{9}$$

for all $x_1, \ldots, x_n$, where $f_{X_1, \ldots, X_n}$ is their joint density.

Independent and identically distributed (i.i.d.) random variables: The random variable

$$X = \begin{cases} 0, & \text{if tails,} \\ 1, & \text{if heads} \end{cases} \tag{10}$$

describes the experiment we perform when tossing a coin once. The probability of the coin landing heads up is an unknown number $0 < p < 1$,

$$\Pr(X = 1) = p.$$

Let's say we choose to toss the coin n times, this gives the random variables

$$X_1, \ldots, X_n,$$

all defined similarly to the random variable X in (10). Since the second toss is not influenced by the outcome of the first toss, the third is not influenced by the second, and so on (this is an assumption), the random variables $X_1, \ldots, X_n$ are *independent.* Moreover, since it is the same coin we are tossing, it is reasonable to assume that the probability p of getting heads does not change from toss to toss, that is

$$\Pr(X_i = 1) = \Pr(X = 1) = p, \quad \text{for } i = 1, \ldots, n.$$

In other words, the random variables $X_1, \ldots, X_n$ are *identically distributed*. Using the independence of $X_1, \ldots, X_n$ and the fact that these are identically distributed, we get

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) = F_X(x_1) \cdots F_X(x_n).$$

In words, the joint distribution of the random variables is equal to the product of the distribution of each single one of them.

2.4. **Expectation.** The expectation of a random variable X is its theoretical mean. Here is an example that should make clear what this means. Let X be a random variable taking its values in $\{0, 1, 2\}$, with distribution,

$$\Pr(X = 0) = \frac{1}{8}, \qquad \Pr(X = 1) = \frac{1}{4}, \qquad \Pr(X = 2) = \frac{5}{8}.$$

Suppose $X_1, \ldots, X_n$ are n independent random variables, all with the same distribution as X. Given that this is all we know, what value would we expect the empirical average $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ to take? Convince yourself of the following

$$\bar{X}_n = 0 \times \frac{\#\{i \colon X_i = 0\}}{n} + 1 \times \frac{\#\{i \colon X_i = 1\}}{n} + 2 \times \frac{\#\{i \colon X_i = 2\}}{n} = \sum_{x=0}^{2} x \frac{\#\{i \colon X_i = x\}}{n},$$

where $\#\{i \colon X_i = x\}$ = the number of i such that $X_i = x$. In this expression for the empirical average $\bar{X}_n$, it certainly seems reasonable that

$$\frac{\#\{i \colon X_i = x\}}{n} \approx \Pr(X = x), \quad \text{for } x = 0, 1, 2,$$

particularly if the sample size n is sufficiently large. This means that $\bar{X}_n$ ought to be close to

$$\sum_{x=0}^{2} x \Pr(X = x) = 0 \times \frac{1}{8} + 1 \times \frac{1}{4} + 2 \times \frac{5}{8} = \frac{3}{2}.$$

Thus, from what we know about the distribution of X, we would expect $\bar{X}_n$ to be close to $3/2$, in fact $3/2$ is the *expectation* or the *expected value* of X.

Here is a Matlab script where we sample $X_1, \ldots, X_n$ for $n = 100$, and then compute the mean in the two different ways indicated above. Run the scrip a few times and see how the empirical mean 'bounces' around its expected value.

```
n = 100 % the sample size
x = randsample([0,1,2],n,true,[1/8, 1/4, 5/8]);
% the true argument in randsample() means that we
% sample with replacement.
mean(x)
0*sum(x == 0)/n + 1*sum(x == 1)/n + 2*sum(x == 2)/n
```

Definition 2.2. (Expectation). The expectation $\mathrm{E}\,X$ of the random variable X taking its values in $\mathcal{X} \subset \mathbb{R} = (-\infty, \infty)$ is given by

$$\mathrm{E}\,X = \sum_{x \in \mathcal{X}} x f(x),$$

when X is discrete (for example $\mathcal{X} = \{0, 1, 2\}$ or $\mathcal{X} = \{0, 1, 2, \ldots\}$), and has pmf. $f(x) = \Pr(X = x)$; and by

$$\mathrm{E}\, X = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x,$$

when X is continuous and has pdf. $f(x)$.

When it makes the math look nicer, we'll sometimes write $\mathrm{E}\,(X)$, $\mathrm{E}\,[X]$, or even $\mathrm{E}\,\{X\}$ instead of $\mathrm{E}\, X$. Also, when g is a real valued function, the expectation of $g(X)$ is

$$\mathrm{E}\, g(X) = \int_{-\infty}^{\infty} g(x) f(x)\, \mathrm{d}x, \tag{11}$$

where X has pdf. $f(x)$. Replace the integral by a sum when X is discrete.

Let's compute the expectation of some random variables. If X takes its values in $\{0, 1\}$ and $\Pr(X = 1) = p$ (a coin flip), then

$$\mathrm{E}\, X = 0 \times \Pr(X = 0) + 1 \times \Pr(X = 1) = 0 \times (1 - p) + 1 \times p = p.$$

The expectation of a fair coin is therefore $1/2$.

If X is a continuous random variable taking its values in $\mathcal{X} = [a, b]$ with equal probability then X has density

$$f(x) = \begin{cases} 1/(b-a), & \text{for } a \le x \le b, \\ 0, & \text{otherwise,} \end{cases}$$

and we say that X has the uniform distribution in $[a, b]$. Its expectation is

$$\mathrm{E}\, X = \int_a^b x \frac{1}{b-a}\, \mathrm{d}x = \frac{1}{2} \frac{x^2}{b-a}\bigg|_a^b = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{1}{2} \frac{(b-a)(b+a)}{b-a} = \frac{b+a}{2}.$$

If X has the exponential distribution on $\mathcal{X} = [0, \infty)$, then its pdf. $f(x)$ is

$$f(x) = \begin{cases} \theta \exp(-\theta x), & \text{for } x \ge 0, \\ 0, & \text{otherwise,} \end{cases}$$

for some $\theta > 0$, then (please show that)

$$\mathrm{E}\, X = \int_{\mathcal{X}} x f(x)\, \mathrm{d}x = \int_0^{\infty} x\, \theta \exp(-\theta x)\, \mathrm{d}x = \frac{1}{\theta}.$$

The most important expectation to know about (for this course) is the expectation of the normal distribution. If $X \sim \mathrm{N}(\mu, \sigma^2)$, which means that X takes its values in $\mathcal{X} = \mathbb{R} = (-\infty, \infty)$, and has the pdf. $f(x)$ given in (7), then

$$\mathrm{E}\, X = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\}\, \mathrm{d}x = \mu.$$

If X is a random variable with pdf. $f(x)$, then for any interval (or union of intervals) in $\mathbb{R}$, the probability that X is in A is

$$\Pr(X \in A) = \int_A f(x)\, \mathrm{d}x.$$

Let I_A be the indicator function,

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Then, using $g(x) = I_A(x)$ in (11), we have

$$\Pr(X \in A) = \int_A f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} I_A(x) f(x)\,\mathrm{d}x = \mathrm{E}\, I_A(X), \tag{12}$$

so $\mathrm{E}\, I_A(X) = \Pr(X \in A)$. For example, $\mathrm{E}\, I_{(-\infty,x]}(x) = \Pr(X \le x) = F(x)$, where F is the cdf. of X.

2.5. **Variance and covariance.** The variance of a random variable X is the expectation of its squared distance from its expectation. We'll write $\mathrm{Var}\, X$, or $\mathrm{Var}\,(X)$, for the variance of X. Here is the definition,

$$\mathrm{Var}\, X = \mathrm{E}\,(X - \mathrm{E}\,[X])^2.$$

For a continuous random variable X with expectation $\mathrm{E}\, X = \mu$ and pdf. X, its variance is

$$\mathrm{Var}\, X = \mathrm{E}\,(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x.$$

Let's compute the variance of the random variable X with the uniform distribution on $[a, b]$. Recall that $\mathrm{E}\, X = (a + b)/2$, and $f(x) = 1/(b - a)$ on $[a, b]$ and zero elsewhere. Thus,

$$\begin{aligned}\mathrm{Var}\, X &= \int_a^b \big(x - \frac{a+b}{2}\big)^2 \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{3}\big(x - \frac{a+b}{2}\big)^3 \frac{1}{b-a}\Big|_a^b \\ &= \frac{1}{3}\Big\{\big(\frac{b-a}{2}\big)^3 - \big(\frac{a-b}{2}\big)^3\Big\}\frac{1}{b-a} = \frac{1}{24}\Big\{\big(\frac{b-a}{2}\big)^3 + \big(\frac{b-a}{2}\big)^3\Big\}\frac{1}{b-a} = \frac{(b-a)^2}{12}.\end{aligned}$$

Here is some Matlab code where we estimate the mean and the variance of a uniform distribution on $[-1, 1]$. Before you run the code, think about what the empirical mean and the empirical variance ought to be close to.

```
x = -1 + 2*rand(100,1); % sample 100 uniforms on [-1,1]
mean(x) % should be close to zero
var(x) % should be close to 1/3
```

The variance of a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$ is σ^2. Recall that its expectation is $\mathrm{E}\, X = \mu$, so

$$\mathrm{Var}\, X = \mathrm{E}\,(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\,\mathrm{d}x = \sigma^2.$$

The covariance of two random variables X and Y, written $\mathrm{Cov}(X, Y)$, is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\,(X - \mathrm{E}\,[X])(Y - \mathrm{E}\,[Y]).$$

A very common distribution when modelling two dependent random variables (X, Y) is the bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and ρ, it has pdf. $f(x, y)$,

$$\begin{aligned} f(x,y) = {} & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ & \times \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)(x-\mu_Y)}{\sigma_X\sigma_Y}\Big)\Big\}. \end{aligned} \tag{13}$$

Here $\rho \in (-1, 1)$ is called the correlation, $\sigma_X, \sigma_Y > 0$, and $\mu_X, \mu_Y \in \mathbb{R}$, and

$$\mathrm{E}\, X = \mu_X, \quad \mathrm{E}\, Y = \mu_Y, \quad \mathrm{Var} X = \sigma_X^2, \quad \mathrm{Var} Y = \sigma_Y^2,$$

while

$$\mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y.$$

The correlation of two random variables X and Y is $\mathrm{Cov}(X, Y)/\sqrt{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}$. For the (X, Y) with pdf. $f(x, y)$ given in (13), the correlation is ρ, for

$$\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = \rho.$$

A simple way of simulating from the bivariate normal distribution with parameter values you choose, is the following. Simulate two independent standard normal random variables $Z_1 \sim \mathrm{N}(0, 1)$ and $Z_2 \sim \mathrm{N}(0, 1)$. Set

$$\begin{aligned} X &= \sigma_X Z_1 + \mu_X, \\ Y &= \sigma_Y(\rho Z_1 + \sqrt{1-\rho^2}Z_2) + \mu_Y. \end{aligned}$$

Then (X, Y) has the joint pdf. $f(x, y)$ given in (13). Here is a Matlab script where we simulate $n = 1000$ independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. (the pairs are independent, not the X_i, Y_i in each pair). Run the script a few times and vary the value of $\rho \in (-1, 1)$ (this is the `rho` in the script).

```
n = 1000;
muX = 0; muY = 0;
sigmaX = 1; sigmaY = 1;
rho = 0.54321;

Z1 = normrnd(0,1,[1,n]);
Z2 = normrnd(0,1,[1,n]);

X = sigmaX*Z1 + muX;
Y = sigmaY*(rho*Z1 + sqrt(1 - rho^2)*Z2) + muY;

scatter(X,Y)
```

2.6. **Properties of expectation and variance.** Suppose that $X_1, \ldots, X_n$ are random variables with joint pdf. $f(x_1, \ldots, x_n)$, and let $g(x_1, \ldots, x_n)$ be a real valued function, thus $g\colon \mathbb{R}^n \to \mathbb{R}$. The expectation of $g(X_1, \ldots, X_n)$ is then

$$\mathrm{E}\, g(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n. \tag{14}$$

If some of these X_is are discrete, the associated integrals are replaced with sums.

If we have two random variables X and Y, whose joint pdf. is $f_{X,Y}(x,y)$. Then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y,$$

is the *marginal* pdf. of X, while $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x$ is the marginal pdf. of Y. Recall also that pdf.'s integrate to 1, so

$$1 = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = 1.$$

Proposition 2.3. *Let $X_1,\dots,X_n$ be random variables, and let $a_1,\dots,a_n$ and b be constants (i.e. not random variables, just some numbers), then*

$$\mathrm{E}\,(a_1X_1 + \cdots + a_nX_n + b) = a_1\mathrm{E}\,(X_1) + \cdots + a_n\mathrm{E}\,(X_n) + b.$$

Proof. For $n = 2$ in class. □

From this proposition it follows that for a random variable X and a constant a

$$\mathrm{Var}\,X = \mathrm{E}\,[X^2] - (\mathrm{E}\,[X])^2, \quad \text{and} \quad \mathrm{Var}\,(aX) = a^2\,\mathrm{Var}(X).$$

Importantly, if $X_1,\dots,X_n$ are i.i.d. random variables, so that they have the same expectation $\mu = \mathrm{E}\,X_1 = \cdots = \mathrm{E}\,X_n$, then Proposition 2.3

$$\mathrm{E}\,\bar{X}_n = \mathrm{E}\,\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\,(X_i) = \mu.$$

We say that the empirical average is unbiased for μ. More on this soon!

Proposition 2.4. *Let X and Y be random variables. Then*

- *If X and Y are independent, then* $\mathrm{Cov}(X,Y) = 0$*;*
- *For constants a, b, c*

$$\mathrm{Var}(aX + bY + c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y).$$

Proof. In class or as homework. □

Let $X_1,\dots,X_n$ be i.i.d. random variables with variance σ^2. We can use Proposition 2.4 to show that

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{\sigma^2}{n}.$$

---

3. Lecture 3, September 7, 2020

See Wooldridge (2019, C-4b p. 725) for a short introduction to maximum likelihood estimation.

Let $X_1, \ldots, X_n$ be some data from a distribution with pdf of pmf $f_\theta(x)$. Here, θ is an unknonwn parameter, or an unknown vector of parameters, that we want to use the data to say something about. It is not obvious how we should use $X_1, \ldots, X_n$ to say something about θ, in other words, it is not obvious how we should construct an estimator, say $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$, that estimates θ.

Maximum likelihood estimation provides a procedure for deriving estimators in problems where one is given a statistical model with some unknown parameter. Let $f_\theta^{\text{joint}}(x_1, \ldots, x_n)$ be the joint pdf or pmf of $X_1, \ldots, X_n$. The *likelihood function* is

$$L_n(\theta) = f_\theta^{\text{joint}}(x_1, \ldots, x_n)$$

When the data $X_1, \ldots, X_n$ are independent – which we will almost always assume – then the likelihood function is

$$L_n(\theta) = f_\theta^{\text{joint}}(x_1, \ldots, x_n) = f_\theta(x_1) \cdots f_\theta(x_n),$$

The likelihood function is a function of θ, when the data is held constant. This means that for different samples of data, you'll get different likelihood functions. The *maximum likelihood estimator*, which we denote by $\widehat{\theta}_n$, is the maximiser of $L_n(\theta)$. That $\widehat{\theta}_n$ maximises $L_n(\theta)$ means that

$$L_n(\widehat{\theta}_n) \geq L_n(\theta) \quad \text{for all } \theta.$$

Since products are difficult to work with, we instead work with the *log-likelihood function.* It is simply the natural logarithm of $L_n(\theta)$, that is

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i),$$

where we in the last equality assume that the data are independent. From now on, we assume that $X_1, \ldots, X_n$ are independent from $f_\theta(x)$. The maximiser of the log-likelihood function $\ell_n(\theta)$ is also the maximiser of the likelihood function $L_n(\theta)$. As said, the likelihood function will change from sample to sample, and so will the log-likelihood function. When deriving estimators it is therefore natural to consider the log-likelihood function as a random variable (but we still write $\ell_n(\theta)$), that is

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

Example 3.1. Let $f_\theta(x) = \theta x^{\theta-1}$ for $x \in [0,1]$, and $f(x) = 0$ for x outside of $[0,1]$, where $\theta > 0$. Suppose that $X_1, \ldots, X_n$ are i.i.d. with pdf $f(x)$. The log of $f_\theta(x)$ is

$$\log f_\theta(x) = \log(\theta) + (\theta - 1)\log x.$$

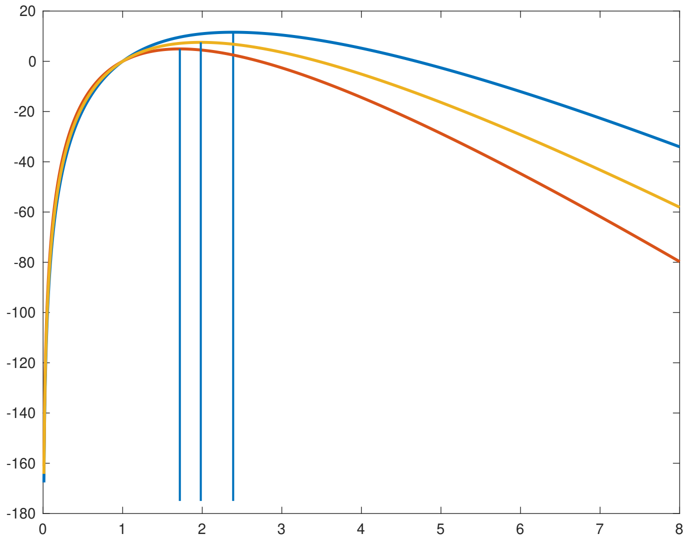

FIGURE 3. The log-likelihood function for three different samples $X_1, \ldots, X_{40}$ from the distribution with density $f_\theta(x) = \theta x^{\theta-1}$ for $x \in [0,1]$, and $f(x) = 0$ for x outside of $[0,1]$, where $\theta > 0$. The vertical lines indicates the different maxima of the functions.

Then

$$\begin{aligned}\ell_n(\theta) &= \sum_{i=1}^n \log f_\theta(X_i) = \sum_{i=1}^n \{\log(\theta) + (\theta-1)\log X_i\} \\ &= n\log\theta + (\theta-1)\sum_{i=1}^n \log X_i.\end{aligned}$$

To find the maximum of $\ell_n(\theta)$ we differentiate with respect to θ and set the derivative equal to zero,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \log X_i = 0. \tag{15}$$

To check that the solution to this equation is indeed a global maximum, we can perform a second derivative test,

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\theta) = -\frac{n}{\theta^2} < 0,$$

for all θ. This means that the solution to $\mathrm{d}\ell_n(\theta)/\mathrm{d}\theta = 0$ is a global maximum, in other words, the function $\ell_n(\theta)$ is everywhere concave. From (15) we see that the maximiser of $\ell_n(\theta)$ is

$$\widehat{\theta}_n = -\frac{n}{\sum_{i=1}^n \log X_i}.$$

This is the *maximum likelihood estimator* (MLE) in this problem.

In Figure 3 I have plotted the log-likelihood function for three simulated samples of size $n = 40$ from the density $f_\theta(x)$. The vertical lines indicates the maxima of the three functions. Notice how the log-likelihood function changes from sample to sample, and consequently, so does the maximum likelihood estimate. Here is the Matlab-script I used to simulate the data and make the figure.

```
theta = 2.34
n = 40;
for sims=1:3
    u = rand(n,1); % random uniform rv's on [0,1]
    x = u.^(1/theta);
    theta_seq = linspace(0.01,8,10^3)
    % the log-likelihood function
    ll_n = n*log(theta_seq) + (theta_seq - 1)*sum(log(x));
    plot(theta_seq,ll_n,'LineWidth', 2)
    theta_hat = -n/sum(log(x))
    line([theta_hat,theta_hat],[-175,max(ll_n)],'LineWidth', 1.414)
    hold on
end
saveas(gcf,"~/your_path/loglik3.eps","epsc");
```

---

4. Lecture 4, September 14, 2020

See Wooldridge (2019, C-3a p. 721) for consistency of estimators, convergence in probability, and the Law of large numbers. What Wooldridge (2019) calls Property PLIM.1 and PLIM.2 will be covered in Lecture 5.

Recall that a sequence of real numbers $(x_n)_{n\geq 1} = (x_1, x_2, x_3, \ldots)$ is said to converge to a number x if for any given $\varepsilon > 0$ we can find a number $N \geq 1$ such that

$$|x_n - x| < \varepsilon, \quad \text{for all } n \geq N.$$

Here is an example: The sequence $x_n = 1/n$ converges to zero. Suppose we are given $\varepsilon = 1/100$, then we can counter with $N = 101$, for certainly

$$|x_n - x| = |1/n| < 1/100 = \varepsilon, \quad \text{for all } n \geq 101.$$

We can also formulate this as follows: That x_n converges to x as $n \to \infty$ means that we can find an $N \geq 1$ such that the set

$$\{n \geq N \colon |x_n - x| \geq \varepsilon\} = \emptyset.$$

Convergence in probability concerns sequences of random variables, say $(X_n)_{n\geq 1} = (X_1, X_2, X_3, \ldots)$, and 'translates' the notion of convergence to a probabilistic statement. Suppose we want to show that X_n converges to a in a probabilistic sense. Instead of asking for an $N \geq 1$ such that $|X_n - a| < \varepsilon$ for all $n \geq N$, we instead ask for an $N \geq 1$ such that the probability of some X_n for $n \geq N$ being more than ε away from a can be made arbitrarily small. Here is the definition.

Definition 4.1. A sequence of random variables $(X_n)_{n\geq 1}$ converges in probability to a constant a if for any given $\varepsilon > 0$

$$\Pr(|X_n - a| \geq \varepsilon) \to 0, \quad \text{as } n \to \infty,$$

we write

$$X_n \xrightarrow{p} a, \quad \text{as } n \to \infty,$$

to indicate convergence in probability of X_n to a.

Another way to say this is: $(X_n)_{n\geq 1}$ converges in probability to a if for any given $\varepsilon > 0$ and $\delta > 0$, we can find $N \geq 1$ such that

$$\Pr(|X_n - a| \geq \varepsilon) < \delta, \quad \text{for all } n \geq N.$$

The typical sequences of random variables that we will meet in this course are sequences of estimators. Say you want to estimate the mean μ of normal distribution. You sample $X_1, \ldots, X_n$ and form the empirical mean $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ and use this as your estimator. Now, for increasing sample size, $(\bar{X}_n)_{n\geq 1}$ is a sequence of random variables, and you want to prove that $\bar{X}_n$ gets close to μ as the sample size n increases.

A very useful inequality when trying to prove that a given sequence of random variables converges in probability to something is Chebyshev's inequality.

Lemma 4.2. (Chebyshev's inequality). *Let X be a random variable with expectation* $\mathrm{E}\, X = \mu$ *and variance* $\mathrm{Var}\, X = \sigma^2$. *Then for any given* $\varepsilon > 0$

$$\Pr(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Recall from the definition $\mathrm{Var}\, X = \mathrm{E}\,(X-\mu)^2$. Assume that X has pdf. $f(x)$ and recall that $f(x) \geq 0$.

$$\begin{aligned}
\sigma^2 = \mathrm{Var}\, X = \mathrm{E}\,(X-\mu)^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&= \int_{|x-\mu|\geq\varepsilon} (x-\mu)^2 f(x)\,\mathrm{d}x + \int_{|x-\mu|<\varepsilon} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&\geq \int_{|x-\mu|\geq\varepsilon} (x-\mu)^2 f(x)\,\mathrm{d}x \geq \int_{|x-\mu|\geq\varepsilon} \varepsilon^2 f(x)\,\mathrm{d}x \\
&= \varepsilon^2 \int_{|x-\mu|\geq\varepsilon} f(x)\,\mathrm{d}x = \varepsilon^2 \Pr(|X-\mu| \geq \varepsilon),
\end{aligned}$$

where in the last equality we use eq. (12). □

Theorem 4.3. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with expectation μ and variance σ^2, and let $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ the empirical mean. Then*

$$\bar{X}_n \xrightarrow{p} \mu, \quad \text{as } n \to \infty.$$

Proof. Use Chebyshev's inequality. In class or as homework. □

---

5. Lecture 5, September 21, 2020

Relevant reading is Wooldridge (2019) Sections C-3a and C3-b, pp. 721–724.

Recall that a sequence of random variables $(X_n)_{n\geq 1}$ converges in probability to a constant a if for any $\varepsilon > 0$,

$$\Pr(|X_n - a| \geq \varepsilon) \to 0,$$

or, equivalently, if

$$\Pr(|X_n - a| < \varepsilon) \to 1,$$

as $n \to \infty$. Why are these two equivalent?

A function $g(x)$ is continuous at x if for any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$|x - y| < \delta \quad \text{implies} \quad |g(x) - g(y)| < \varepsilon.$$

A function that is continuous at every point x in some interval of the real line, is continuous on this interval. One can think of a continuous function as a "function that you can graph without lifting your pencil from the paper" (Wooldridge, 2019, p. 722). Here are some continuous function: $g(x) = a+bx$ for constant a and b, $g(x) = x^2$, $g(x) = 1/x$, $g(x) = \sqrt{x}$, $g(x) = log(x)$, $g(x) = \exp(x)$. Also, a composition of continuous function is a continuous function. For example, the function $h(x) = \exp(a + bx)$ is continuous. The next lemma is called Property PLIM.1 in Wooldridge (2019, p. 722). Note that Wooldridge (2019) writes $\text{plim}(X_n) = a$ when I write $X_n \to_p a$.

Lemma 5.1. (Prop. PLIM.1) *Let X_n be a sequence of rv's and a a constant. If $X_n \to_p a$ and $g(x)$ is a continuous function, then $g(X_n) \to_p g(a)$.*

Proof. Since $g(x)$ is continuous we know that for any $\varepsilon > 0$ we can find $\delta > 0$ such that $|x - a| < \delta$ implies $|g(x) - g(a)| < \varepsilon$. In terms of events, this means that for any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$\{|X_n - a| < \delta\} \subset \{|g(X_n) - g(a)| < \varepsilon\}.$$

By Proposition 1.3(f), this means that

$$\Pr(|X_n - a| < \delta) \leq \Pr(|g(X_n) - g(a)| < \varepsilon).$$

By Proposition 1.3(b) $\Pr(|g(X_n) - g(a)| < \varepsilon) \leq 1$, and by assumption $\Pr(|X_n - a| < \delta) \to 1$, so since $\Pr(|g(X_n) - g(a)| < \varepsilon)$ is squeezed in between, $\Pr(|g(X_n) - g(a)| < \varepsilon) \to 1$. □

The next lemma is called Property PLIM.2 in Wooldridge (2019, p. 723).

Lemma 5.2. (Prop. PLIM.2) *Assume that $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are sequences of random variables, that a and b are constants, and that $X_n \to_p a$ and $Y_n \to_p b$. Then*

(i) $X_n + Y_n \to_p a + b$;
(ii) $X_n Y_n \to_p ab$;
(iii) $X_n / Y_n \to_p a/b$ *provided* $b \neq 0$.

Remark 5.3. A sequence of numbers $(b_n)_{n\geq 1}$ that converges to a constant b in the sense discussed at the start of this lecture, also converges in probability to b. Thus, if $X_n \to_p a$, and $b_n \to b$, then it follows from Lemma 5.2(i) that $X_n + b_n \to_p a+b$; from Lemma 5.2(ii) that $X_n b_n \to_p ab$; and from Lemma 5.2(iii) that $X_n/b_n \to_p a/b$, provided $b \neq 0$. For example, if $X_n \to_p a$, then $X_n/n \to_p 0$.

Proof. (of Lemma 5.2). We will prove (i), the rest is in Homework 5. We want to prove that for any given $\varepsilon > 0$,

$$\Pr(|X_n + Y_n - (a+b)|\geq \varepsilon) \to 0.$$

We have

$$|X_n + Y_n - (a+b)|= |(X_n - a) + (Y_n - b)|\leq |X_n - a|+|Y_n - b|,$$

by the triangle inequality. In terms of events, this means that

$$\{|X_n + Y_n - (a+b)|\geq \varepsilon\} \subset \{|X_n - a|+|Y_n - b|\geq \varepsilon\},$$

so by Prop. 1.3(f) it is sufficient to show that $\Pr(|X_n - a|+|Y_n - b|\geq \varepsilon) \to 0$, since

$$0 \leq \Pr(|X_n + Y_n - (a+b)|\geq \varepsilon) \leq \Pr(|X_n - a|+|Y_n - b|\geq \varepsilon).$$

Given $\varepsilon > 0$ and for $n = 1, 2, \ldots$, defined the event

$$\begin{aligned} A_n &= \{|X_n - a|+|Y_n - b|\geq \varepsilon\} \\ B_n &= \{|Y_n - b|\geq \varepsilon/2\}, \end{aligned}$$

so that $B_n^c = \{|Y_n - b|< \varepsilon/2\}$. We now want to show that $\Pr(A_n) \to 0$. By the Law of total probability (see hw1 Ex. 4(b)), and using that $\Pr(A_n \mid B_n)\Pr(B_n) \leq \Pr(B_n)$, we get

$$\begin{aligned} \Pr(A_n) &= \Pr(A_n \cap B_n) + \Pr(A_n \cap B_n^c) = \Pr(A_n \mid B_n)\Pr(B_n) + \Pr(A_n \cap B_n^c) \\ &\leq \Pr(B_n) + \Pr(A_n \cap B_n^c). \end{aligned}$$

Here $\Pr(B_n) = \Pr(|Y_n - b|\geq \varepsilon/2) \to 0$ by assumption, so we now only need to show that $\Pr(A_n \cap B_n^c) \to 0$. But when $|Y_n - b|\geq \varepsilon/2$, which it is in the intersection $A_n \cap B_n^c$, then $|X_n - a|+|Y_n - b|\leq |X_n - a|+\varepsilon/2$. Therefore,

$$\begin{aligned} A_n \cap B_n^c &= \{|X_n - a|+|Y_n - b|\geq \varepsilon\} \cap \{|Y_n - b|< \varepsilon/2\} \\ &\subset \{|X_n - a|+\varepsilon/2 \geq \varepsilon\} \cap \{|Y_n - b|< \varepsilon/2\} \\ &\subset \{|X_n - a|+\varepsilon/2 \geq \varepsilon\}, \end{aligned}$$

where for the last inequality we use that for any two event A and B, $A \cap B \subset A$ (and also $A \cap B \subset B$), draw a Venn diagram. But this shows that $\Pr(A_n \cap B_n^c) \leq \Pr(|X_n - a|\geq \varepsilon/2)$, and in summary

$$0 \leq \Pr(|X_n - a|+|Y_n - b|\geq \varepsilon) \leq \Pr(|X_n - a|\geq \varepsilon/2) + \Pr(|Y_n - b|\geq \varepsilon/2),$$

where the sum on the right tends to zero by assumption. $\square$

Convergence in distribution. We now turn to another form of convergence of random variables. Let $X_1, X_2, \ldots$ be a sequence of random variables, and $F_1, F_2, \ldots$ the corresponding sequence of cumulative distributions functions, that is $F_i(x) = \Pr(X_i \leq x)$ for $i = 1, 2, \ldots$. Let X be a random variable with cdf $F(x) = \Pr(X \leq x)$. We say that X_n converges in distribution to X, and write

$$X_n \xrightarrow{d} X,$$

if, as $n \to \infty$,

$$F_n(x) \to F(x),$$

for all points x at which $F(x)$ is continuous.

Example 5.4. (Don't spend much time on this example. I included it to show that a limiting distribution is not always normal. See hw5 Ex. 4 for another non-normal limit distribution.) For each $n = 1, 2, \ldots$ let X_n be a random variable that takes its values in

$$\{1/n, 2/n, \ldots, (n-1)/n, 1\},$$

with equal probability, i.e. $\Pr(X_n = j/n) = 1/n$ for $j = 1, \ldots, n$. Recall that if X is a random variable with the uniform distribution on $[0, 1]$, then its cdf is $F(x) = x$ for $x \in [0, 1]$, $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 1$ (see hw2 Ex. 7). For $x \in \{1/n, 2/n, \ldots, (n-1)/n, 1\}$, the cdf of X_n is

$$F_n(x) = \sum_{j=1}^{\lfloor nx \rfloor} \frac{1}{n} = \frac{\lfloor nx \rfloor}{n},$$

where $\lfloor y \rfloor = \max\{m \in \{0, 1, 2, \ldots\} \mid m \leq y\}$ is called the floor function. Let $\mathrm{frac}(y)$ be the fraction part of y, for example $\mathrm{frac}(2.34) = 0.34$, so that $\mathrm{frac}(y) = y - \lfloor y \rfloor$, for example $0.34 = \mathrm{frac}(2.34) = 2.34 - \lfloor 2.34 \rfloor = 2.34 - 2$. This means that, $0 \leq \mathrm{frac}(y) < 1$ for all y. We can write

$$F_n(x) = \frac{\lfloor nx \rfloor}{n} = \frac{nx + \mathrm{frac}(nx)}{n} = x + \frac{\mathrm{frac}(nx)}{n} \to x,$$

as $n \to \infty$, which means $X_n \to_d X$, where X is a uniform random variable on $[0, 1]$.

The central limit theorem. There are several central limit theorems, so the 'the' in the header is not that precise, but I'll use it anyways. We have seen that $X_n \to_d X$ means that $F_n(x) \to F(x)$, with $X_n \sim F_n$ for $n = 1, 2, \ldots$, and $X \sim F$. The central limit theorem (CLT) concerns cases where the limiting cdf F of the sequence of cdf's $(F_n)_{n \geq 1}$ is that of a normal distribution. Since the normal distribution, and in particular the *standard* normal distribution appears so often, we reserve special symbols for its pdf and its cdf: If $Z \sim \mathrm{N}(0, 1)$, then its pdf is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad z \in (-\infty, \infty),$$

and its cdf is

$$\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \phi(x) \, \mathrm{d}x.$$

As an exercise, suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$ so that X has cdf

$$F_{\mu,\sigma}(x) = \int_{\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}\,\mathrm{d}y.$$

Show that

$$F_{\mu,\sigma}(x) = \Phi\big(\frac{x-\mu}{\sigma}\big),$$

and that

$$\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$$

The next theorem can be found in Wooldridge (2019, p. 724).

Theorem 5.5. (Central limit theorem). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation $\mathrm{E}[X_1] = \mu$ and variance $\mathrm{Var}(X_1) = \sigma^2$, and set $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Define*

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then

$$Z_n \xrightarrow{d} Z, \quad \text{where } Z \sim \mathrm{N}(0,1).$$

In other words, if $Z_n \sim F_n(z)$ for $n = 1, 2, \ldots$, then

$$F_n(z) \to \Phi(z), \quad \text{for each } z.$$

Why does it matter? Notice that the only assumptions we make about the $X_1, X_2, \ldots$ in the theorem are that they are independent, identically distributed, and that they have an expectation and a variance. We do not say anything more about their distribution. For example, if we were asked to compute the probabilities $\Pr(X_1 \leq x)$, or $\Pr(Z_{23} \leq z)$, we would be at loss. The CLT, however, tells us that for n sufficiently large (what 'sufficiently large' means can often be checked by way of simulations)

$$\Pr(Z_n \leq z) \approx \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\,\mathrm{d}x,$$

and the integral on the right we *can* compute. Here is a command computing $\Phi(1.96)$ in Matlab

```
normcdf(1.96,0,1)
```

So if you want to compute $\Pr(-1.96 \leq Z_n \leq 1.96)$, use that (see hw1, Ex. 11)

$$\Pr(-1.96 \leq Z_n \leq 1.96) = \Pr(Z_n \leq 1.96) - \Pr(Z_n \leq -1.96), \approx \Phi(1.96) - \Phi(-1.96)$$

for n sufficiently large, then go to Matlab and type

```
normcdf(1.96,0,1) - normcdf(-1.96,0,1)
```

to get 0.95. The inverse of the normal cdf $\Phi^{-1}(p)$ also exists in Matlab, for example

```
norminv(0.975,0,1)
```

returns 1.96, and `norminv(normcdf(1.96,0,1),0,1)` also returns 1.96, etc.

Family	Father	Mother	Gender	Height	Kids
1	78.5	67	M	73.2	4
1	78.5	67	F	69.2	4
1	78.5	67	F	69	4
1	78.5	67	F	69	4
2	75.5	66.5	M	73.5	4
2	75.5	66.5	M	72.5	4
2	75.5	66.5	F	65.5	4
2	75.5	66.5	F	65.5	4
3	75	64	M	71	2
3	75	64	F	68	2
4	75	64	M	70.5	5
4	75	64	M	68.5	5
4	75	64	F	67	5
4	75	64	F	64.5	5
4	75	64	F	63	5
5	75	58.5	M	72	6
5	75	58.5	M	69	6

Table 1. A subset of the Galton height data set. The full dataset contains 898 rows, and 197 families. Heights are given in inches: one inch = 2.54 cm. The full data set is available from many websites, including the Harvard dataverse website, where I found it.

6. Lecture 6, September 28, 2020

Relevant reading is Wooldridge (2019) Sections 2.1–2.6 on regression analysis.

Suppose we have a dataset with two measurements on n units (individuals, families, firms, stocks, schools, or whatever unit you like to think of). Table 1 gives the first few rows of a famous dataset collected by Francis Galton in England in the 1880's. I found it on the Harvard dataverse website, you can also read more about, and look at photos of the original dataset here. The dataset contains the heights (in inches) of mothers, fathers, and their children. In all there are 898 rows (thus 898 mother, father, child triplets), and 197 families. The full dataset is available on Itslearning in the file `galton.txt`. Since the data include multiple children per family, the variable Family is a family ID variable; the variable Father is the height of the father; the variable Mother is the height of the mother; Gender is gender; Height is height; and Kids is the number of children the family has.

In Figure 4 I have plotted the heights of all $n = 433$ mother–daughter pairs. It is natural to think that the height of your mother influences your height, or, more mathematically speaking, that your height is a function of your mothers height. At the same time, however, there are clearly many other factors that also influences the height of a daughter. A model that captures such a 'non-perfect' or non-deterministic relation between the height

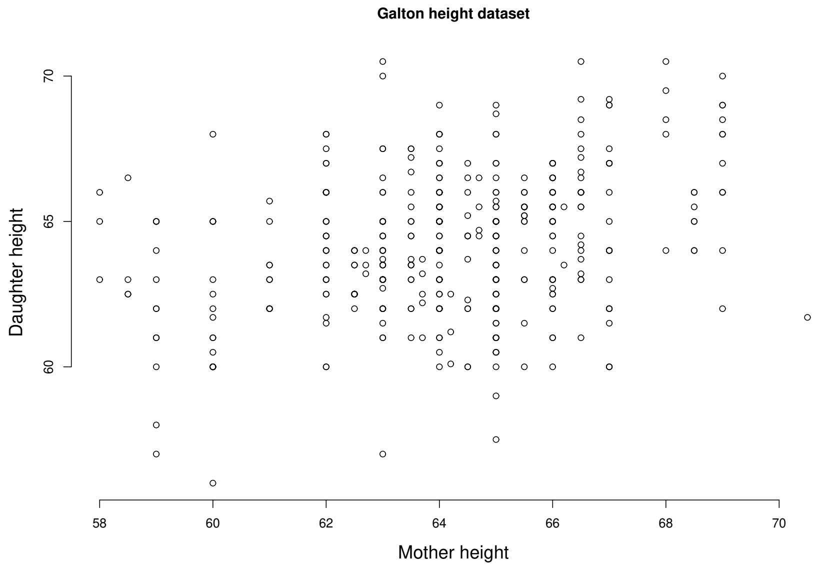

FIGURE 4. All mother–daughter pairs in the Galton height datasets.

of mothers and the height of their daughters is the following:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{16}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with $\mathrm{E}\left[\varepsilon_1\right] = 0$ and $\mathrm{Var}(\varepsilon_1) = \sigma^2$; and

$$\begin{aligned} Y_i &= \text{Height of daughter } i, \\ x_i &= \text{Height of mother } i, \end{aligned}$$

for $i = 1, \ldots, n$, where we $x_1, \ldots x_n$ are fixed numbers; and β_0 and β_1 are unknown regression coefficients, or parameters. The x_i's are variously called independent variables, covariates, features, and surely others things as well. The Y_i's are called the dependent variables, the outcome, and also other things.

Notice that each Y_i is a function of the random variable ε_i, hence itself a random variable. It follows from Proposition 2.3 that for each i,

$$\mathrm{E}\left[Y_i\right] = \mathrm{E}\left[\beta_0 + \beta_1 x_i + \varepsilon_i\right] = \beta_0 + \beta_1 x_i + \mathrm{E}\left[\varepsilon_i\right] = \beta_0 + \beta_1 x_i,$$

thus we see that the intercept β_0 is the *expected* value of Y_i when $x_i = 0$, and that the slope β_1 is the *expected* increase in Y_i with a one unit increase in x_i.

In terms of the $n = 433$ mother–daughter pairs in Figure 4, let's have a close look at the assumptions we are making when postulating the model in (16) (these are what Wooldridge (2019, pp. 40–) calls SLR.1–SLR.5. Make drawings, and make sure you understand these.

(i) **Linear in parameters**. The function $y(x) = \beta_0 + \beta_1 x$ is on average the correct way to describe the relation between the height a daughter and the height of her mother;

(ii) **Random sampling**. The noise terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent, meaning that the height of the ith daughter does not tell us anything about the height of the jth daughter ($i \neq j$);
(iii) **Sample variation in the explanatory variable**. The $x_1, \ldots, x_n$ are not all the same. If all the mothers were of the same height, mothers height couldn't possibly create variation in the height of the daughters.
(iv) **Zero conditional mean**. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ have expectation zero no matter the value of x_i. The other factors influencing the height of a daughter cancels out on average, no matter the height of the mother.
(v) **Homoskedasticity**. Is the $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i assumption: The variance of the error terms are the same no matter where you are on the x-axis. How much the height of a daughter might deviate from her expected height $\mathrm{E}\,[\text{Height of daughter}] = \beta_0 + \beta_1 \text{Height of mother}$, is the same not matter the height of the mother.

As we go along, we will see when these assumptions are important. The parameters β_0, β_1, and σ^2 are in most, if not all, real world applications unknown, so we need to estimate these from the data

$$(x_1, Y_1), \ldots, (x_n, Y_n).$$

Since the model in (16) postulates that the relation between the x_i's and the Y_i's is a line, we can ask for the line that best fits the data. What is natural to consider a good line, is a line that makes the distance between each Y_i and $\beta_0 + \beta_1 x_i$ small. We don't care whether our point $\beta_0 + \beta_1 x_i$ is below or above Y_i, so we square the distances $Y_i - (\beta_0 + \beta_1 x_i)$. Make a drawing! The least squares estimators are the minimisers of the function

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

We denote the least squares estimators by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, thus

$$g(\widehat{\beta}_0, \widehat{\beta}_1) \leq g(\beta_0, \beta_1), \quad \text{for all } \beta_0, \beta_1.$$

To find these we take the partial derivatives with respect to β_0 and β_1, and set these expressions equal to zero. This gives two equations in two unknowns,

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} g(\beta_0, \beta_1) &= -2n(\bar{Y}_n - \beta_0 - \beta_1 \bar{x}_n) = 0, \\
\frac{\partial}{\partial \beta_1} g(\beta_0, \beta_1) &= \sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) - \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = 0,
\end{aligned}$$

where $\bar{Y}_n = (1/n)\sum_{i=1}^{n} Y_i$ and $\bar{x}_n = (1/n)\sum_{i=1}^{n} x_i$. The solution to these equations are

$$\widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{x}_n, \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}.$$

These are the least squares estimators. Each time we see an estimator, or construct a new estimator, there we should ask several questions, some of which are: (i) What is the expectation if the estimator? Is it biased or unbiased; (ii) What is the variance of the estimator? (iii) Is the estimator consistent? (iv) What is the distribution, or approximate distribution of the estimator? In hw6 Ex. 1 you are asked to find the expectation and variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Fitted values. What we refer to as the estimated line, or fitted line of a regression, is the line

$$(x, \widehat{\beta}_0 + \widehat{\beta}_1 x).$$

The quantities

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \quad \text{for } i = 1, \dots, n,$$

are referred to as the fitted values, or predicted values. They are our estimates of $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$ for $i = 1, \dots, n$.

Residuals. When we are to draw conclusion about the real world using a the model in (16) (or any statistical model, for that matter), our conclusions are only valid as long as our assumptions are valid. It is therefore important to try to assess whether the assumptions hold. Plotting the residuals can help. The residuals from fitting a regression are the

$$u_i = Y_i - \widehat{Y}_i, \quad \text{for } i = 1, \dots, n.$$

Since $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$, the residuals can be viewed as estimates of the error terms. And since the error terms have expectation zero and constant variance (the Homoskedasticity assumption), so should the $u_1, \dots, u_n$ if the model is any good.

Let's look at the residuals in a case where we know that the Assumptions (i)–(v) hold, and in a case where one or more of them is broken. To do this, we simulate data.

Example 6.1. (Residuals when assumptions hold). To simulate data from the model in (16) we need to make an extra assumption about the error terms $\varepsilon_1, \dots, \varepsilon_n$. In addition to the assumptions already made, we will assume that they are normally distributed.

```
cd("your path")
n = 400;
beta0 = -0.543; beta1 = 2.345;
x = linspace(0,1,n);
sigma2 = 1.234
eps = normrnd(0,sqrt(sigma2),1,n);
y = beta0 + beta1.*x + eps;
scatter(x,y)
beta1hat = sum((x - mean(x)).*y)/sum((x - mean(x)).^2)
beta0hat = mean(y) - beta1hat*mean(x)

yhat = beta0hat + beta1hat.*x; % The fitted values
u = y - yhat; % The residuals

scatter(x,u)
hold on
plot([0,1],[0,0],"Color","g","Linewidth",2)
xlabel("x");ylabel("Residuals")
```

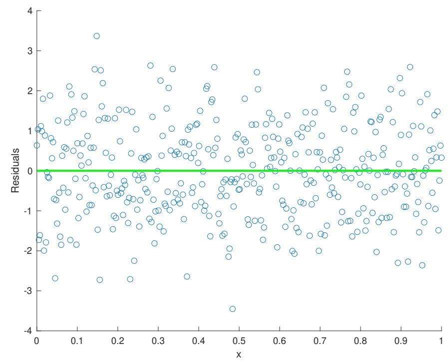

FIGURE 5. The nice residuals simulated in Example 6.1.

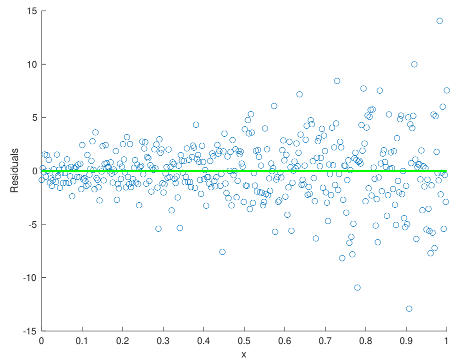

FIGURE 6. The bad residuals simulated in Example 6.2.

```
saveas(gcf,"your path/niceresid.eps","epsc")
```

The residuals are plotted in Figure 5. Notice how they are centered around zero for all values of x, and how their spread around zero is the about the same for all values of x.

Example 6.2. (RESIDUALS WHEN ASSUMPTION (V) IS BROKEN). In this simulation example we are going to break Assumption (v), namely the assumption of homoskedasticity. We do this by taking the variance to be a function of the independent variable. Here, we'll take

$$\sigma^2(x) = 1.234 \exp(3x).$$

This is the only modification we do to the Matlab script in Example 6.1. The residuals from one such simulation are plotted in Figure 6. Notice how the spread of the residuals around the green line at zero (their expectation) increases as x increases.

7. Lecture 7, October 5, 2020

In addition to the relevant reading for last week, Wooldridge (2019) Sections 2.1–2.6, which is still relevant, please look at Wooldridge (2019) Sections 3.1–3.3, and Math Refreshers (Appendices) C-5 and C-6 on interval estimation and confidence intervals and hypothesis testing, respectively.

The normal distribution (Parts of this is repetition from Lecture 5). We write $X \sim \mathrm{N}(a, b^2)$ when X is a normally distributed random with a normal expectation $\mathrm{E}[X] = a$ and variance $\mathrm{Var}(X) = b^2$. The pdf of X is

$$f_{a,b}(x) = \frac{1}{\sqrt{2\pi}b} \exp\big\{-\frac{(x-a)^2}{2b^2}\big\}, \quad x \in (-\infty, \infty),$$

and its cdf is

$$F_{a,b}(x) = \int_{-\infty}^{x} f_{a,b}(y)\,\mathrm{d}y,$$

for all x. If $Z \sim \mathrm{N}(0,1)$, we say that Z has the *standard* normal distribution, and reserve special symbols for its pdf and cdf,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad \text{and} \quad \Phi(z) = \int_{-\infty}^{z} \phi(y)\,\mathrm{d}y.$$

So in terms of the $f_{a,b}(x)$ just above, $\phi(x) = f_{0,1}(x)$.

Lemma 7.1. *If $X \sim \mathrm{N}(a, b^2)$, then*

$$\frac{X-a}{b} \sim \mathrm{N}(0,1).$$

Proof. We use the symbols just introduced.

$$\begin{aligned}
\Pr\big(\frac{X-a}{b} \le z\big) &= \Pr(X \le bz + a) = F_{a,b}(bz+a) \\
&= \int_{-\infty}^{bz+a} \frac{1}{\sqrt{2\pi}b} \exp\big\{-\frac{1}{2}\big(\frac{y-a}{b}\big)^2\big\}\,\mathrm{d}y = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}b} \exp(-w^2/2)\,b\,\mathrm{d}w \\
&= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-w^2/2)\,\mathrm{d}w = \Phi(z),
\end{aligned}$$

where we used the substitution $w = (y-a)/b$ so that $\mathrm{d}x = b\,\mathrm{d}w$. This shows that the cdf of $(X-a)/b$ is $\Phi(z)$, which means that $(X-a)/b$ is a standard normal random variable. □

Here is a lemma that we will not prove, but use very often.

Lemma 7.2. *Let $X_1, \ldots, X_n$ be independent random variables with distributions $N(a_i, b_i^2)$ for $i = 1, \ldots, n$, i.e. $X_1 \sim N(a_1, b_1^2)$, and so on. Let $\gamma_1, \ldots, \gamma_n$ and η be constants (not random variables), then*

$$\sum_{i=1}^{n} \gamma_i X_i + \eta \sim \mathrm{N}\big(\sum_{i=1}^{n} \gamma_i a_i + \eta, \sum_{i=1}^{n} \gamma_i^2 b_i^2\big).$$

Proof. Not part of the curriculum. □

As an exercise, you can try to deduce Lemma 7.1 from Lemma 7.2.

Example 7.3. Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$, then Lemma 7.2 entails that

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

and if we combine this with Lemma 7.1, we get that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathrm{N}(0,1).$$

Example 7.4. Consider the regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$, and $x_1, \ldots, x_n$ are constants, not all equal. Then Lemma 7.2 combined with hw6 Ex. 1(g) gives that the least squares estimator

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \sim \mathrm{N}\big(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\big).$$

The *standard error* of $\widehat{\beta}_1$ is the square root of its variance, we write

$$\mathrm{se}(\widehat{\beta}_1) = \frac{\sigma}{\{\sum_{i=1}^n (x_i - \bar{x}_n)^2\}^{1/2}}$$

From Lemma 7.1 we have that

$$\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \sim \mathrm{N}(0,1). \tag{17}$$

so that

$$\Pr\big(\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq z\big) = \Phi(z).$$

This is a key result when we construct confidence intervals for β_1, and derive tests for hypotheses about β_1.

If in the model of Example 7.4 the errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathrm{E}[\varepsilon_1] = 0$ and $\mathrm{Var}(\varepsilon_1) = \sigma^2$, but we do *not* assume that they are normally distributed, then the least squares estimator $\widehat{\beta}_1$ does *not* have a normal distribution. In particular, (17) is not true. It does, however, have expectation $\mathrm{E}[\widehat{\beta}_1] = \beta_1$ and variance $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2 / \sum_{i=1}(x - \bar{x}_n)^2$.

Fortunately, the relation in (17) is approximately true, thanks to the a central limit theorem. The next theorem is not in itself part of the curriculum, but you should know about what it says about the approximate distribution of least squares estimators.

Theorem 7.5. (The Lindeberg–Lévy central limit theorem). *Let $X_1, \ldots, X_n$ be independent random variables with expectation 0 and variances $\sigma_1^2, \ldots, \sigma_n^2$, and set $B_n^2 = \sum_{i=1}^n \sigma_i^2$. Then*

$$\frac{1}{B_n}\sum_{i=1}^n X_i \xrightarrow{d} \mathrm{N}(0,1),$$

as $n \to \infty$, provided the Lindeberg condition is satisfied. This conditions says that, for any $\delta > 0$,

$$\frac{1}{B_n^2} \sum_{i=1}^{n} \mathrm{E}\left[X_i^2 I\{|X_i| \geq \delta B_n\}\right] \to 0,$$

as $n \to \infty$.

Proof. The proof is not part of the curriculum. If you are interested, you can look at a proof I wrote for a course I taught last year (Stoltenberg, 2019). □

Notice that the random variables $X_1, \ldots, X_n$ in this theorem are required to be independent, but not identically distributed. This is the difference between this theorem and Theorem 5.5, where the random variables are i.i.d. (independent *and* identically distributed).

Consider again the regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathrm{E}[\varepsilon_1] = 0$ and $\mathrm{Var}(\varepsilon_1) = \sigma^2$. If we had assumed that the $\varepsilon_1, \ldots, \varepsilon_n$ were normal, Lemma 7.2 would give that $\widehat{\beta}_1 \sim \mathrm{N}(0, \sigma^2/\{\sum_{i=1}^n (x_i - \bar{x}_n)^2\})$, where $\widehat{\beta}_1$ is the least squares estimator. But we do *not* assume that the errors are normal! To proceed with inference on β_1 (confidence intervals, tests, etc.) we would like to use a central limit theorem to approximate the distribution of $\widehat{\beta}_1$. If we define

$$a_i = \frac{x_i - \bar{x}_n}{\sum_{i=1}(x_i - \bar{x}_n)^2}, \quad \text{for } i = 1, \ldots, n,$$

we can write

$$\widehat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i \varepsilon_i,$$

Thus, the difference $\widehat{\beta}_1 - \beta_1$ is equal to the sum of the random variables,

$$a_1\varepsilon_1, \ldots, a_n\varepsilon_n.$$

These are independent, but since

$$\mathrm{Var}(a_i\varepsilon_i) = a_i^2\sigma^2 = \frac{(x_i - \bar{x}_n)^2\sigma^2}{\{\sum_{j=1}^n (x_j - \bar{x}_n)^2\}^2},$$

they are *not* identically distributed (this variance depends on the index i). This is why we need the Lindeberg-Lévy central limit theorem, and not merely Theorem 5.5, to get a an approximation to the distribution of the least squares estimator $\widehat{\beta}_1$. Let

$$B_n^2 = \sum_{i=1}^{n} \mathrm{Var}(a_i\varepsilon_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma^2}{\sum_{j=1}^n (x_j - \bar{x}_n)^2}.$$

Then Theorem 7.5 says that

$$\frac{1}{B_n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{d} \mathrm{N}(0, 1), \tag{18}$$

as $n \to \infty$, provided the Lindeberg condition holds. You do not need to worry about this condition in this course, but it does not hurt to know that here, the condition is satisfied as long as

$$\frac{\max_{i \leq n} |x_i - \bar{x}_n|}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \to 0,$$

as $n \to \infty$, meaning that no covariate value is too different from the others, and thus none of the random variables $a_i \varepsilon_i$ has a variance that is much larger than the variance of the others.

The result in (18) is extremely important, for it is this result that allows us to use the approximation

$$\Pr\big(\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq z\big) \approx \Phi(z), \tag{19}$$

when n is large, where $\mathrm{se}(\widehat{\beta}_1) = \sigma/\{\sum_{i=1}^n (x_i - \bar{x}_n)^2)\}^{1/2}$. As we will soon see (in Lecture 8, perhaps), this approximation is still valid when we replace $\mathrm{se}(\widehat{\beta}_1)$ by an estimator, $\widehat{\sigma}_n/\{\sum_{i=1}^n (x_i - \bar{x}_n)^2\}$, where $\widehat{\sigma}_n$ is an estimator of σ.

The approximation in (19) allows us to build confidence intervals and perform tests for β_1. For example, let

$$z_{\alpha/2} = \Phi^{-1}(\alpha/2), \quad \text{and} \quad z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

where α is the your chosen significance level, and

$$z_{\alpha/2} = -z_{1-\alpha/2},$$

by the symmetry of the normal distribution. Often $\alpha = 0.05$, in which case

$$z_{0.025} = -1.96 = \Phi^{-1}(0.025) = \Phi^{-1}(\alpha/2), \quad \text{and} \quad z_{0.975} = 1.96 = \Phi^{-1}(0.975) = \Phi^{-1}(1-\alpha/2).$$

You can find these numbers by typing `norminv(0.025,0,1)` and `norminv(0.975,0,1)` in Matlab. Then (using hw 1, Ex. 11),

$$\begin{aligned}\Pr\big(-z_{1-\alpha/2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq z_{1-\alpha/2}\big) &= \Pr\big(\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq z_{1-\alpha/2}\big) - \Pr\big(\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq -z_{1-\alpha/2}\big)\\ &\approx \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) = \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2})\\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha,\end{aligned}$$

which is equal to 0.95 when $\alpha = 0.05$. This means that

$$\Pr\big(-z_{1-\alpha/2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}(\widehat{\beta}_1)} \leq z_{1-\alpha/2}\big) = \Pr\big(\widehat{\beta}_1 + z_{\alpha/2}\,\mathrm{se}(\widehat{\beta}_1) \leq \beta_1 \leq \widehat{\beta}_1 + z_{1-\alpha/2}\,\mathrm{se}(\widehat{\beta}_1)\big) \approx 1-\alpha,$$

So if σ is known – which it rarely, if ever, is – then

$$[\widehat{\beta}_1 + z_{\alpha/2}\,\mathrm{se}(\widehat{\beta}_1), \widehat{\beta}_1 + z_{1-\alpha/2}\,\mathrm{se}(\widehat{\beta}_1)]$$

is an *approximate* $(1-\alpha) \times 100$ percent confidence interval for β_1. In actual applications, you will need to estimate σ, but the approximate inequalities above do still hold, so

$$\big[\widehat{\beta}_1 + z_{\alpha/2}\,\widehat{\mathrm{se}}(\widehat{\beta}_1), \widehat{\beta}_1 + z_{1-\alpha/2}\,\widehat{\mathrm{se}}(\widehat{\beta}_1)\big],$$

is also an approximate $(1-\alpha)\times 100$ percent confidence interval for β_1, where $\widehat{\mathrm{se}}(\widehat{\beta}_1)$ is our estimator of $\mathrm{se}(\widehat{\beta}_1)$, that is

$$\mathrm{se}(\widehat{\beta}_1) = \frac{\sigma}{\{\sum_{i=1}^n (x_i-\bar{x}_n)^2\}^{1/2}}, \quad \text{and} \quad \widehat{\mathrm{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}_n}{\{\sum_{i=1}^n (x_i-\bar{x}_n)^2\}^{1/2}}$$

where $\widehat{\sigma}_n$ is a consistent estimator for σ, typically $\widehat{\sigma}_n$ is the square root of

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2,$$

or its unbiased version, see hw6, Ex. 2(e).

As another, but closely related, example of the use of the approximation in (19), say you want to test the hypotheses,

$$H_0\colon \beta_1 = 0 \quad \text{vs.} \quad H_A\colon \beta_1 \neq 0.$$

A natural test is to reject H_0 if

$$\widehat{\beta}_1 \le -c_n \quad \text{or} \quad \widehat{\beta}_1 \ge c_n, \tag{20}$$

for some $c_n > 0$, where c_n is chosen so that

$$\Pr(\text{Type I error}) \approx \alpha,$$

where α is the significance level (that you set!). Then, assuming that H_0 is true (that is, assuming that $\beta_1 = 0$),

$$\begin{aligned}
\Pr(\text{Type I error}) &= \Pr_{H_0}(\widehat{\beta}_1 \le -c_n \text{ or } \widehat{\beta}_1 \ge c_n) \\
&= \Pr_{H_0}(\widehat{\beta}_1 \le -c_n) + \{1 - \Pr_{H_0}(\widehat{\beta}_1 \le c_n)\} \\
&= \Pr_{H_0}\big(\frac{\widehat{\beta}_1}{\mathrm{se}(\widehat{\beta}_1)} \le -\frac{c_n}{\mathrm{se}(\widehat{\beta}_1)}\big) + \{1 - \Pr_{H_0}\big(\frac{\widehat{\beta}_1}{\mathrm{se}(\widehat{\beta}_1)} \le \frac{c_n}{\mathrm{se}(\widehat{\beta}_1)}\big)\} \\
&\approx \Phi\big(-\frac{c_n}{\mathrm{se}(\widehat{\beta}_1)}\big) - \{1 - \Phi\big(\frac{c_n}{\mathrm{se}(\widehat{\beta}_1)}\big)\} = \alpha,
\end{aligned}$$

where the approximate inequality stems from (19), and the last equality is true provided

$$c_n = \Phi^{-1}(1-\alpha/2)\,\mathrm{se}(\widehat{\beta}_1).$$

(The comment made above about estimating $\mathrm{se}(\widehat{\beta}_1)$ applies here as well.) This means that the test in (20), with the appropriately chosen c_n, is a test for H_0 at *approximately* the $\alpha \times 100$ percent significance level.

---

8. Lecture 8, October 12, 2020

In this lecture we will study regression models with more than one independent variable. Say we have data

$$(x_{1,1}, x_{1,2}, Y_1), \dots, (x_{n,1}, x_{n,2}, Y_n),$$

from n individuals (schools, firms, stocks, etc.), and we postulate the model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, \quad \text{for } i = 1, \dots, n,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent random variables with expectation zero and variance σ^2, and $x_{1,1}, \dots, x_{n,1}, x_{1,2}, \dots, x_{n,2}$ are independent variables. Relevant reading for this lecture is Sections 3.1–3.4, and Section 4.1–4.5 in Wooldridge (2019). Also take a look at Math Refreshers B4-e, B4-f, and B4-g, on conditional expectation, properties of conditional expectation, and conditional variance, respectively (Wooldridge, 2019, pp. 700–704).

References

Allen, L. J. (2003). *An Introduction to Stochastic Processes with Applications to Biology.* Pearson Prentice Hall, Upper Saddle River, NJ.

Casella, G. and Berger, R. L. (2002). *Statistical Inference. Second Edition.* Duxbury Pacific Grove, CA.

Grimmet, G. and Stirzaker, D. (2001). *Probability and Random Processes. Third Edition.* Oxford University Press, Oxford.

Hammack, R. H. (2020). *Book of Proof. Third Edition.* Creative Commons. Available here: `https://www.people.vcu.edu/~rhammack/BookOfProof/`.

Jacod, J. and Protter, P. (2012). *Probability Essentials.* Springer, New York.

Papineau, D. (2012). *Philosophical devices: Proofs, probabilities, possibilities, and sets.* Oxford University Press, Oxford.

Shiryaev, A. (1996). *Probability. Second Edition.* Springer, New York.

Stoltenberg, E. A. (2019). A moment generating function proof of the central limit theorem. STK4011, Autumn semester 2019, Department of Mathematics, University of Oslo. `https://www.uio.no/studier/emner/matnat/math/STK4090/v20/lindebergclt_mgf.pdf`.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach. Seventh Edition.* Cengage Learning, Boston, MA.

Department of Economics, BI Norwegian Business School
Email address: `emil.a.stoltenberg@bi.no`