18.600: Lecture 23

Conditional probability, order statistics, expectations of sums

Scott Sheffield

MIT

Outline

Outline

Conditional probability densities

Order statistics

Expectations of sums

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.
- We can *define* the conditional probability density of X given that $Y = y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.
- We can *define* the conditional probability density of X given that $Y = y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.
- This amounts to restricting $f(x, y)$ to the line corresponding to the given y value (and dividing by the constant that makes the integral along that line equal to 1).

Conditional distributions

- Let's say X and Y have joint probability density function $f(x,y)$.
- We can *define* the conditional probability density of X given that $Y=y$ by $f_{X|Y=y}(x)=\frac{f(x,y)}{f_Y(y)}$.
- This amounts to restricting $f(x,y)$ to the line corresponding to the given y value (and dividing by the constant that makes the integral along that line equal to 1).
- This definition assumes that $f_Y(y)=\int_{-\infty}^{\infty} f(x,y)dx<\infty$ and $f_Y(y)\neq 0$. This *usually* safe to assume. (It is true for a probability one set of y values, so places where definition doesn't make sense can be ignored).

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?
- Tough question in general.

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?
- Tough question in general.
- Consider conditional law of X given that $Y \in (y - \epsilon, y + \epsilon)$. If this has a limit as $\epsilon \to 0$, we can call *that* the law conditioned on $Y = y$.

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?
- Tough question in general.
- Consider conditional law of X given that $Y \in (y - \epsilon, y + \epsilon)$. If this has a limit as $\epsilon \to 0$, we can call *that* the law conditioned on $Y = y$.
- Precisely, define $F_{X|Y=y}(a) := \lim_{\epsilon \to 0} P\{X \leq a | Y \in (y - \epsilon, y + \epsilon)\}$.

Remarks: conditioning on a probability zero event

- Our standard definition of conditional probability is $P(A|B) = P(AB)/P(B)$.
- Doesn't make sense if $P(B) = 0$. But previous slide defines "probability conditioned on $Y = y$" and $P\{Y = y\} = 0$.
- When can we (somehow) make sense of conditioning on probability zero event?
- Tough question in general.
- Consider conditional law of X given that $Y \in (y - \epsilon, y + \epsilon)$. If this has a limit as $\epsilon \to 0$, we can call *that* the law conditioned on $Y = y$.
- Precisely, define $F_{X|Y=y}(a) := \lim_{\epsilon \to 0} P\{X \le a | Y \in (y - \epsilon, y + \epsilon)\}$.
- Then set $f_{X|Y=y}(a) = F'_{X|Y=y}(a)$. Consistent with definition from previous slide.

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0, 1]$ (zero otherwise).

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0, 1]$ (zero otherwise).
- Let (θ, R) be (X, Y) in polar coordinates. What is $f_{X|\theta=0}(x)$?

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0, 1]$ (zero otherwise).
- Let (θ, R) be (X, Y) in polar coordinates. What is $f_{X|\theta=0}(x)$?
- Answer: $f_{X|\theta=0}(x) = 2x$ if $x \in [0, 1]$ (zero otherwise).

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x, y) : x^2 + y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0, 1]$ (zero otherwise).
- Let (θ, R) be (X, Y) in polar coordinates. What is $f_{X|\theta=0}(x)$?
- Answer: $f_{X|\theta=0}(x) = 2x$ if $x \in [0, 1]$ (zero otherwise).
- Both $\{\theta = 0\}$ and $\{Y = 0\}$ describe the same probability zero event. But our interpretation of what it means to condition on this event is different in these two cases.

A word of caution

- Suppose X and Y are chosen uniformly on the semicircle $\{(x,y): x^2+y^2 \leq 1, x \geq 0\}$. What is $f_{X|Y=0}(x)$?
- Answer: $f_{X|Y=0}(x) = 1$ if $x \in [0,1]$ (zero otherwise).
- Let (θ, R) be (X,Y) in polar coordinates. What is $f_{X|\theta=0}(x)$?
- Answer: $f_{X|\theta=0}(x) = 2x$ if $x \in [0,1]$ (zero otherwise).
- Both $\{\theta = 0\}$ and $\{Y = 0\}$ describe the same probability zero event. But our interpretation of what it means to condition on this event is different in these two cases.
- Conditioning on (X,Y) belonging to a $\theta \in (-\epsilon, \epsilon)$ wedge is very different from conditioning on (X,Y) belonging to a $Y \in (-\epsilon, \epsilon)$ strip.

Outline

Outline

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0, 1]$, independently of each other.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0, 1]^n$.
- What is the probability that the *largest* of the X_i is less than a?

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0, 1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \ldots, X_n\}$, then what is the probability density function of X?

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \dots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \dots, X_n\}$, then what is the probability density function of X?
- Answer: $F_X(a) = \begin{cases} 0 & a < 0 \\ a^n & a \in [0,1] \\ 1 & a > 1 \end{cases}$. And $f_x(a) = F_X'(a) = na^{n-1}$.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \ldots < x_n$, zero otherwise.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^n f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^n f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \dots, Y_n)$ independent of each other?

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \ldots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \ldots, Y_n)$ independent of each other?
- Yes.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?
- Write $p = X_1$. This kind of like choosing a random p and then conditioning on 7 heads and 2 tails.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?
- Write $p = X_1$. This kind of like choosing a random p and then conditioning on 7 heads and 2 tails.
- Answer is beta distribution with parameters $(a, b) = (8, 3)$.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0, 1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?
- Write $p = X_1$. This kind of like choosing a random p and then conditioning on 7 heads and 2 tails.
- Answer is beta distribution with parameters $(a, b) = (8, 3)$.
- Up to a constant, $f(x) = x^7(1 - x)^2$.

Example

- Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables on $[0,1]$.
- Example: say $n = 10$ and condition on X_1 being the third largest of the X_j.
- Given this, what is the conditional probability density function for X_1?
- Write $p = X_1$. This kind of like choosing a random p and then conditioning on 7 heads and 2 tails.
- Answer is beta distribution with parameters $(a, b) = (8, 3)$.
- Up to a constant, $f(x) = x^7(1-x)^2$.
- General beta (a, b) expectation is $a/(a+b) = 8/11$. Mode is $\frac{(a-1)}{(a-1)+(b-1)} = 2/9$.

Outline

Outline

Conditional probability densities

Order statistics

Expectations of sums

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x, y)$ then $E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x,y)$ then $E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$.
- If X and Y have joint probability density function $f(x,y)$ then $E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X + Y] = E[X] + E[Y]$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we have $E[X] = \int_0^\infty P\{X > x\}$, in discrete and continuous settings?

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we have $E[X] = \int_0^\infty P\{X > x\}$, in discrete and continuous settings?
- Yes, can prove with integration by parts or...

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we have $E[X] = \int_0^\infty P\{X > x\}$, in discrete and continuous settings?
- Yes, can prove with integration by parts or...
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we have $E[X] = \int_0^\infty P\{X > x\}$, in discrete and continuous settings?
- Yes, can prove with integration by parts or...
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we have $E[X] = \int_0^\infty P\{X > x\}$, in discrete and continuous settings?
- Yes, can prove with integration by parts or...
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.
- So $E[X] = E[g(Y)] = \int_0^1 g(y)dy$, which is indeed the area under the graph of $1 - F_X$.