ECON-C4100 - Capstone: Econometrics I

Lecture 2B: Statistics recap

Otto Toivanen

Learning outcomes: conditional descriptive statistics

- After this lecture you understand
  1. the meaning of central concepts for conditional descriptive statistics of a variable,
  2. how to characterize the conditional distributions,
  3. how to characterize distributions of more than one variable more generally, and
  4. why conditional descriptive statistics are a first step towards causal analysis.

Learning outcomes: random sampling and estimation of the mean

- By the end of the lecture, you

  5. know what random sampling means.
  6. appreciate the difference between **population** and ***sample***.
  7. understand the concept of **independently and identically distributed**.
  8. understand why the sample mean is (almost) never equal to the population mean, but is correct on average.
  9. know what **an estimator** is.
  10. know what **an estimate** is.
  11. understand the concepts of **bias**, **consistency** and **efficiency** of an estimator.
  12. understand that an estimator is a **random variable**.
  13. why the sample mean is **BLUE**.

2. What are conditional descriptive statistics?

- Conditional descriptive statistics are characterized by the researcher *conditioning* the information on Y on another variable X.
- Simple but important example: conditional mean.

$$\mathbb{E}[Y|X = x]$$

- Conditional descriptive statistics build on the *joint distribution* of two or more variables.
- We will work with the case of two variables.

From joint density to individual density

- How might we get the density function of X in the case of a observing two (discrete) variables X and Y?

$$f_X(x) = \sum_y f_{X,Y}(x, y) \tag{1}$$

- Such a density function is called the *marginal distribution* (of X).
- Notice that the marginal distribution takes into account all values of X irrespective of what value Y takes (or, for all values of Y).

From marginal to conditional distribution

- What if we are interested in what values Y gets, conditional on a given value x of X?
- Then we are interested in a *conditional distribution*, or some function of it.
- The conditional distribution of Y given X = x is defined as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x).} \quad (2)$$

- The conditional distribution is not defined when $f_X(x) = 0$.

Visualizing a joint distribution

- How to visualize your data consisting of two variables?
- A scatter-plot allows you to display all of your data.
- Example: our FLEED analysis sample.
- Let's add age to our analysis.
- FLEED contains variable $syntyv = YoB$.

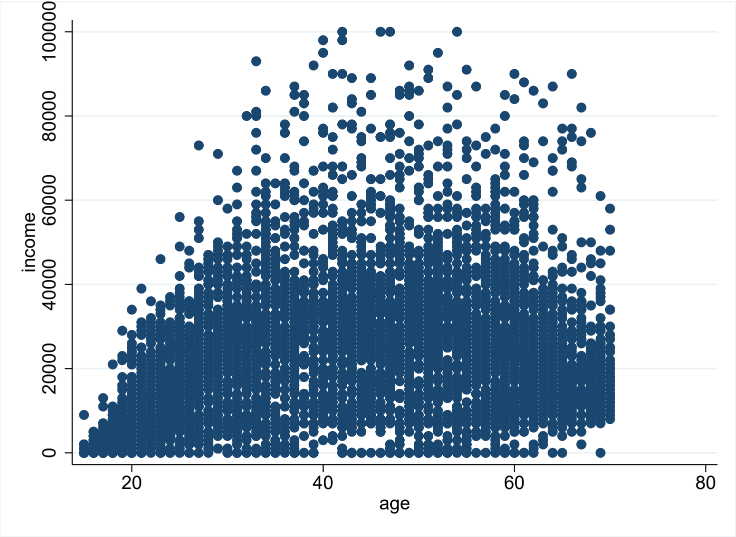
Visualizing a joint distribution

- Let's draw a scatter plot of income as a function of age.

Stata code

```
twoway scatter income age if year == 15 & income != . || ///
lfit income age if year == 15 & income != . , ///
xtitle("age") ///
graphregion(fcolor(white))
graph export "income_age_line.pdf", replace
```

Scatterplot of income and age, analysis sample



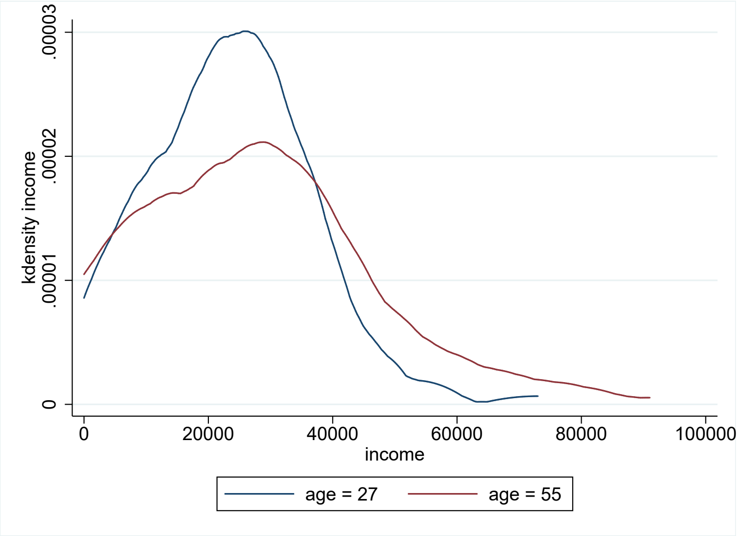
Conditional distributions

- How do the distributions of income at two different ages compare?
- Let's start by comparing two density plots.

Stata code

```
twoway  kdensity income if year == 15 & income != . & age == 27 ||  ///
kdensity income if year == 15 & income != . & age == 55 , ///
xtitle("income")  ///
legend(label(1 "age = 27") label(2 "age = 55")) ///
graphregion(fcolor(white))
graph export "income_distr_age27_age55.pdf", replace
```

Density plot of income for age = 27, age = 55



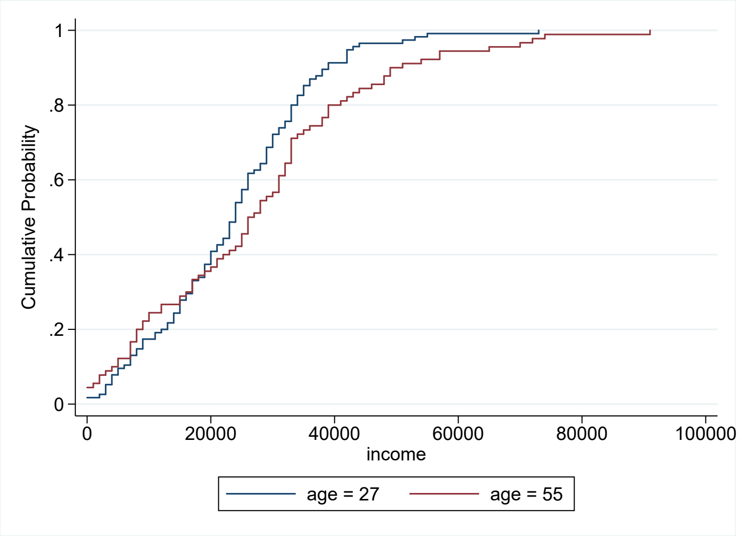
What about the cdfs?

- Just like in the univariate case, the density plot is informative in its own way, the cdf in another way.

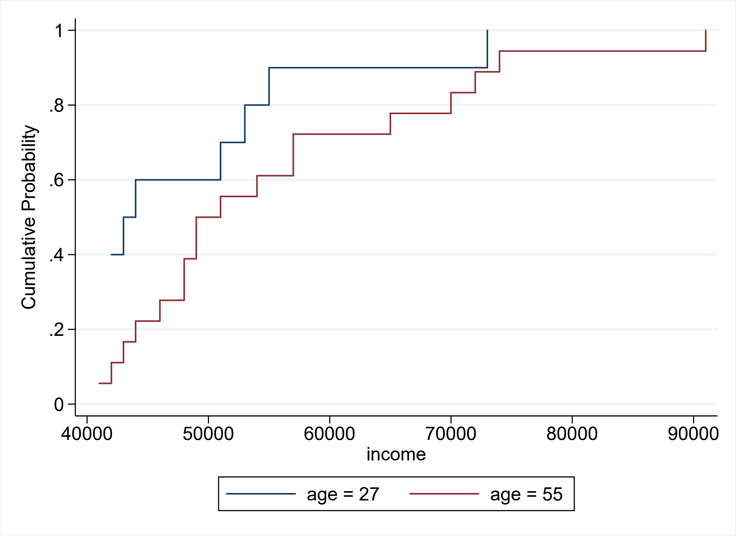
Stata code

```
gen young = .
replace young = 0 if age == 27
replace young = 1 if age == 55
cdfplot income if year == 15 & income != . & young != ., by(young) ///
xtitle("income") ///
legend(label(1 "age = 27") label(2 "age = 55")) ///
graphregion(fcolor(white))
graph export "income_cdf_age27_age55.pdf", replace
```

Cdf's of income for age = 27, age = 55



Cdf's of income for age = 27, age = 55, income > 40 000



Conditional means

1. A key concept in empirical economics is the conditional mean

$$\mathbb{E}[Y|X = x]$$

2. What would these look like in the analysis data on income, if X is age?

Stata code

```
tabstat income if year == 15 & income != ., stat(mean) by(age)
```

Income conditional on age

age	mean income
15	411
20	8 346
27	23 565
30	24 011
40	31 430
50	30 082
55	27 411
60	26 407
70	19 344
Total	23 297

Income conditional on age

- How does income develop with age?
- How much does age increase income in expectation, going from 30 to 40 years?
- Why might the mean income of 50+ be lower than that of those aged 40?
- Aside: at what level of accuracy should we report mean incomes (1 euro, 1 000 euros, ...)?

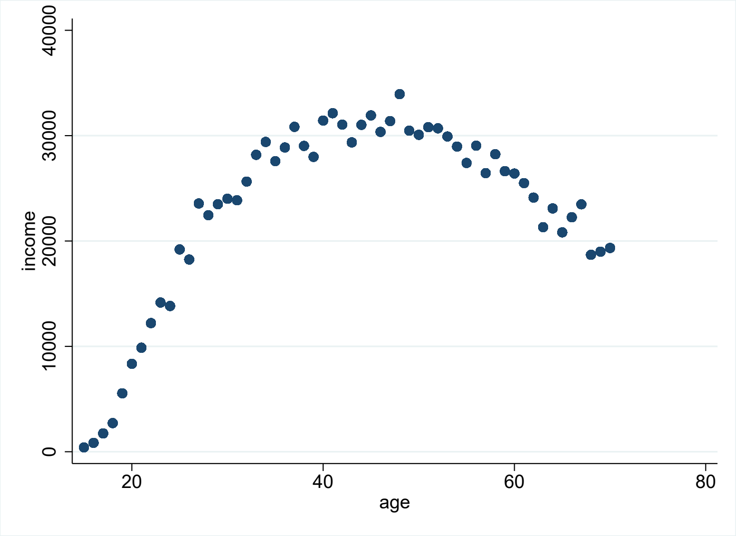
Income conditional on age

- Imagine you wanted to study the causal effect of X on Y. Conditional means allow you to study the correlation of them, forming a first step towards causal analysis.
- Showing a table for all ages in the data leads to a very large table.
- How else could one display the incomes conditional on age?

Stata code

```
bysort age: egen income_age_m = mean(income) if year == 15 & income != .
scatter income_age_m age if year == 15 & income != . & income_age_m != . , ///
xtitle("age") ytitle("income") ///
graphregion(fcolor(white)) \linebreak
graph export "income_age_condmean.pdf", replace
```

Mean income conditional on age



Correlation

- The best known descriptive statistic to characterize how two variables' values are aligned is *correlation*.
- To get to correlation, we need to first define the *covariance*.
- The covariance of Y and X is defined as

$$\begin{aligned} Cov(X,Y) &= \mathbb{E}[X-\mathbb{E}(X)]\,\mathbb{E}[Y-\mathbb{E}(Y)] \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i)(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i), \end{aligned} \tag{3}$$

- And the correlation of Y and X as

$$Cor(X,Y) = cov(X,Y)/[sd(X)sd(Y)]. \tag{4}$$

2. Random sampling and estimation of the mean

- Example of random sampling: Finland conducted an experiment on basic income in 2017 - 2018. (see Verho, J., Hämäläinen, K. & Kanninen, O. (2021). Removing welfare traps: Employment responses in the finnish basic income experiment. *American Economic Journal: Economic Policy, forthcoming*).
- For the purposes of the basic income study, a random sample from the target population was drawn.
- The important numbers for the random sampling were:
  1. 175 000 individuals in the (target) population.
  2. 2 000 individuals drawn from this population into the treatment group.

Population and sample

- Population = those units that we are interested in (N).
- Sample = those units that we select out of the population, i.e., a subset of the population (n).

Random sampling

- Random sampling = each object in the population has the same probability of being selected into the sample.
- Two key requirements: Each subject is
  1. Independently distributed = any two objects are not informative about each other.
     - Y and X are independent iff $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.
  2. Identically distributed = before being chosen, each object is equal in expectation.
     - Y and X are identically distributed iff $F_X(x) = F_Y(x)$.
- *Random variable* = numerical summary of a random outcome.

Random sampling - class room experiment

- We collected data on the height and gender of the students of this course.
- I treat those students who answered as the *population* and take random samples from it.
- Questions to be solved prior to commencing:
  1. How many students to include in the sample?
  2. How to choose them?

Random sampling - class room experiment

- In our data $N = 45$.
- I chose $n = 3, 5, 9, 15$.
- In standard random sampling, I would have chosen n once and selected one random sample of size n.
- Now I draw as many samples of size n as I can as long as I only sample each individual only once.

Random sampling - class room experiment

- Let's first have a look at the population data.
- Notice that in usual circumstances we would not have access to these data.
- It is the mean of the population height that we try to estimate through our random sample(s).

Mean	sd	Median
177.0	10.8	179

Estimating the mean of a population

- *Estimator* = some function of sample data.
- *Estimate* = the numerical value of the estimator, *given a particular sample*.
- Notice that the sample mean $(= \bar{Y})$ is not the same as the population mean, but a natural *estimator* of it.
- Consequently, 177.0 is not our *estimate* of the sample mean (it **is** the population mean, i.e., the target of our estimation); we are about to study several such estimates.

Estimating the mean of a population

- Two questions.

  1. What are the properties of $\bar{Y}$?
  2. Why use $\bar{Y}$ instead of some other estimator?

Properties of $\bar{Y}$

- $\bar{Y}$ is a random variable.
- Its properties are determined by the *sampling distribution*.
- The individual observations used to calculate $\bar{Y}$ were chosen (iid) randomly.
- What happens to $\bar{Y}$ if you take another random sample (of size n)?
- The *sampling distribution* = the distribution of $\bar{Y}$ over **all possible samples of size** n.
- Example: All possible samples of size 9 (=all possible combinations of 9 students) from the population of students that submitted their height information.

Estimates of $\bar{Y}$ based on $n = 9$

Group					Population
1	2	3	4	5	Population mean
170.3	175.8	177.8	179.1	181.1	177.0

Properties of $\bar{Y}$

- Sampling distribution:
  1. all the values $\bar{Y}$ can take
  2. the probability of each of these values.
- The mean and variance of $\bar{Y}$ are the mean and variance of its sampling distribution.

Properties of an estimator of μ_Y

- NOTE: at the risk of confusion, I use the more general notation of $\hat{\mu}_Y$ for the estimator on this slide, not $\bar{Y}$.
- The reason is that these properties apply generally.
- Let $\hat{\mu}_Y$ be an estimator of μ_Y.
  1. The **bias** of $\hat{\mu}_Y = \mathbb{E}(\hat{\mu}_Y) - \mu_Y$.
  2. $\hat{\mu}_Y$ is **unbiased estimator** of μ_Y if $\mathbb{E}[\hat{\mu}_Y] = \mu_Y$.
  3. $\hat{\mu}_Y$ is a **consistent** estimate of μ_Y if $\hat{\mu}_Y \rightarrow \mu_Y$ when $n \rightarrow \infty$.
  4. let $\tilde{\mu}_Y$ be another unbiased estimator of μ_Y. Then $\hat{\mu}_Y$ is more **efficient** than $\tilde{\mu}_Y$ if $var(\hat{\mu}_Y) < var(\tilde{\mu}_Y)$.
- These properties of an estimator are **generic**, i.e., they apply to any estimator.

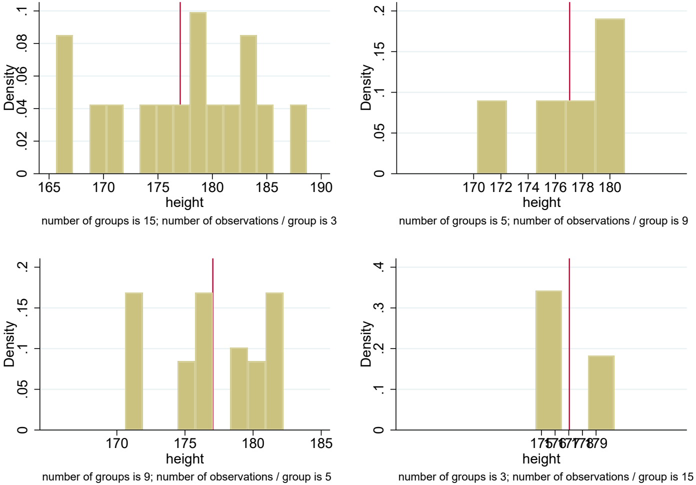
Properties of $\bar{Y}$

- Due to the *Law of Large Numbers*, $\bar{Y}$ is both *unbiased* and *consistent*.
- LLN requires that the sample is iid.

Estimating the mean - class room experiment

- Let's demonstrate consistency and the effect of sample size with our height data.
- On the next slide are graphs of the distributions of our estimates of $\bar{Y}$ using different n.
- The vertical red line is the "truth", i.e., the population mean of 177.0.

Estimating the mean - class room experiment



Estimating the mean - class room experiment

- In each graph, each estimate is unbiased (= on average, they are correct).
- As we increase the sample size from the upper left graph ($n = 3$) to the lower right corner ($n = 15$) the $\bar{Y}$ - estimates get closer to the population mean.
- This is what consistency means.

Properties of $\bar{Y}$

- How precise is $\bar{Y}$, and how does this depend on n?
- In other words, how large is the variance of $\bar{Y}$?
- *The Central Limit Theorem* gives the answer.
- Hint: look at how close the estimates $\bar{Y}$ are to the population mean as we vary sample size n in the graph above.

Central Limit Theorem

- The CLT
  1. is about the distribution of the *estimate* of the mean.
  2. applies *no matter* what the distribution of the underlying variable Y is.
- Examples: coin tosses (binary), age (only positive/integer)

How the mean becomes normally distributed with large enough samples

- Example: Draws from a Poisson distribution with an increasing n.
- Demonstration of how the distribution develops courtesy of Richard Hennigan.

Properties of $\bar{Y}$

- The CLT shows that the following hold:
- Suppose
  1. the sample is iid.
  2. $\mathbb{E}[Y] = \mu_Y$.
  3. $var(Y) = \sigma_Y^2 < \infty$

Properties of $\bar{Y}$

- Then, as $n \to \infty$, the distribution of $\bar{Y}$ becomes arbitrarily well approximated by the normal distribution $N(\mu_Y, \sigma^2_{\bar{Y}})$.

  Notice that the variance of this normal distribution is decreasing in n.
- Then, as $n \to \infty$, the distribution of

$$\frac{\bar{Y} - \mu_Y}{\sigma^2_Y}$$

  becomes arbitrarily well approximated by the standard normal distribution $N(0, 1)$.

$\bar{Y}$ as a least squares estimator

- $\bar{Y}$ minimizes the *sum of squared residuals*:

$$min_m \sum_{i=1}^{N} (y_i - m)^2 \tag{5}$$

- $\bar{Y}$ has smaller variance than all other unbiased linear estimators.
- $\rightarrow$ $\bar{Y}$ is more efficient than other (linear) estimators.
- $\bar{Y}$ is **B**est **L**inear **U**nbiased **E**stimator (BLUE).

Testing the mean

- Imagine you want to test whether the $\bar{Y}$ you estimated is different from some value Y_0.
- The *t-statistic* is given by

$$t = (\bar{Y} - Y_0)/\hat{\sigma}_Y \quad (6)$$

where $\hat{\sigma}_Y = s_Y/\sqrt{n}$ is the estimated standard error of $\bar{Y}$.

- The distribution of t is appr. standard normal (why?).
- Notice how the denominator depends on n.
- This is the reason why a larger sample is beneficial in terms of testing hypotheses, i.e., statistical significance.