University of Exeter

Applied Econometrics

Topic 2: Exploratory Data Analysis

Jingya Zeng

University of Exeter

Table of contents

University of Exeter

Exploratory data analysis

University of Exeter

Goals of exploratory data analysis

Term “exploratory data analysis”, as opposed to “confirmatory data analysis”, is due to John Tukey (1977, 1980). Its aims are:

- uncover the underlying structure of the data;
- test underlying assumptions for statistical procedures;
- detect outliers and anomalies;
- develop parsimonious models.
- Explore your data open-mindedly!

Disclaimer: It doesn’t mean that you should build your models solely from exploring the data.

Methods

EDA mainly relies on non-rigorous, graphical methods, and on tables and lists.

These tools are often accompanied by other methods:

- statistical test procedures
- data reduction techniques
- fitting a function to the data (e.g. linear or quadratic)

The goal is to gain a better insight into your data, without imposing many assumptions on structure.

University of Exeter

Univariate distributions

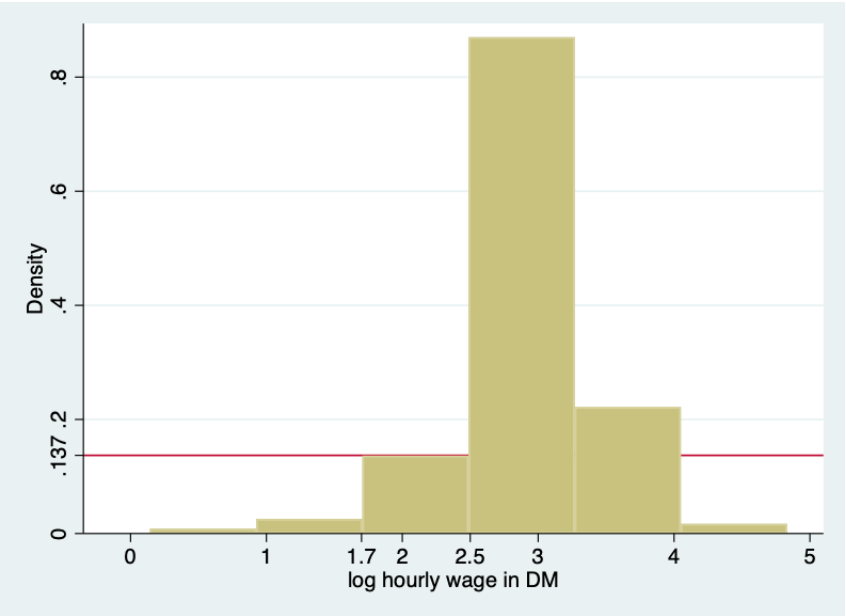
One of the first things to look in a new data set are the distributions of the main variables.

`summarize` gives a first impression (example with the pen01aL3 data)

```
. summarize lnw, detail

                    log hourly wage in DM
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.018646       .1431778
 5%     2.194448       1.018646
10%     2.345427        1.24179       Obs                 150
25%     2.631166       1.619084       Sum of wgt.         150

50%     2.969492                      Mean           2.901829
                       Largest        Std. dev.      .5269156
75%     3.199535       3.818812
90%     3.406059       3.968189       Variance       .2776401
95%     3.579777       4.055201       Skewness      -1.150735
99%     4.055201       4.831729       Kurtosis       9.284479
```

Histograms

```
histogram lnw, bin(6) yline(0.137) xlabel(1.7 2.5, add) ylabel(0.137, add)
```



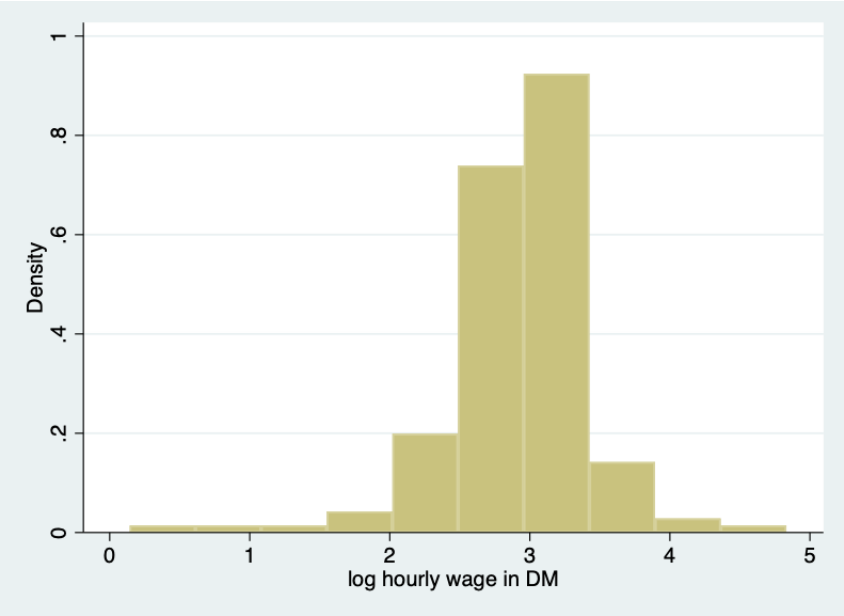
In a histogram, the area of a bar is equal to the (observed) probability that an observation falls into the interval.

In principle, the bars could be of different widths, but this makes a histogram harder to interpret.

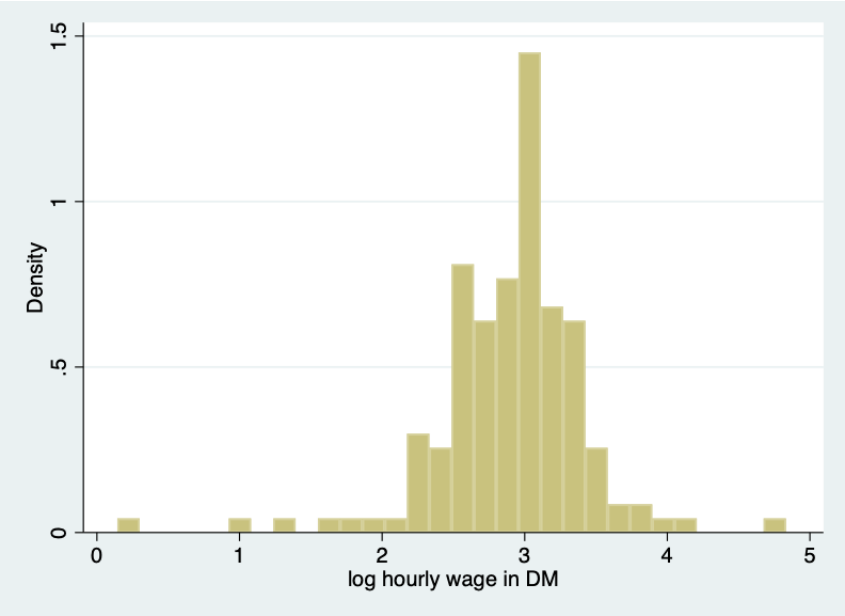
Histograms

Histograms are very sensitive to the width of bins:

histogram lnw, width(0.4688552)



histogram lnw, width(0.15628504)

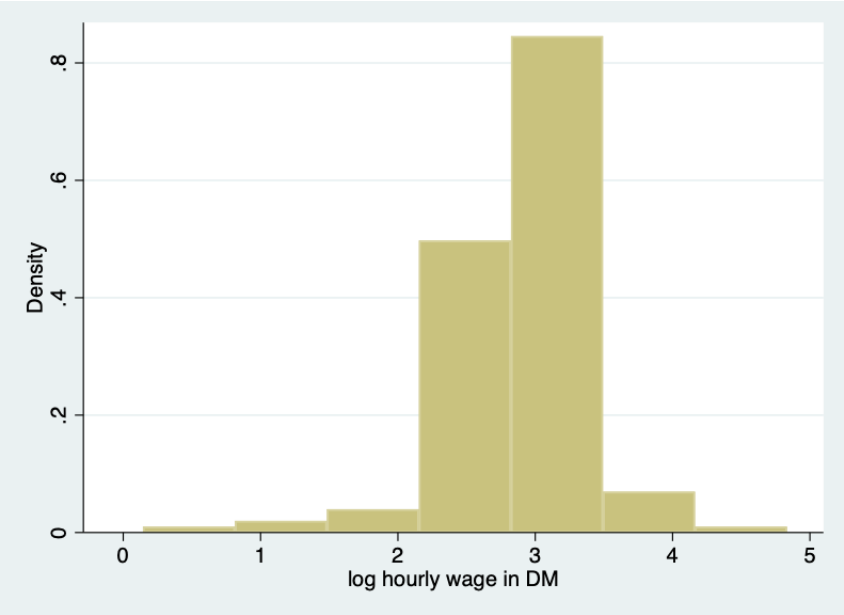


University of Exeter

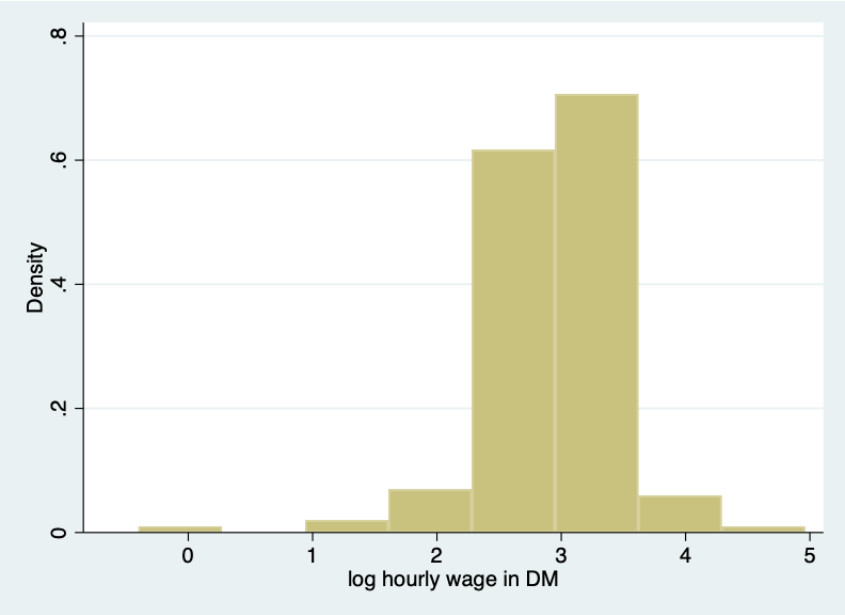
Histograms

For a given bin width, they are sensitive to the start point:

`histogram lnw, start(.1431778) width(.67)`



`histogram lnw, start(-0.4) width(.67)`



University of Exeter

Histograms

To summarise, histograms are:

- dependent on the bin width;
- dependent on the start point;
- not smooth.

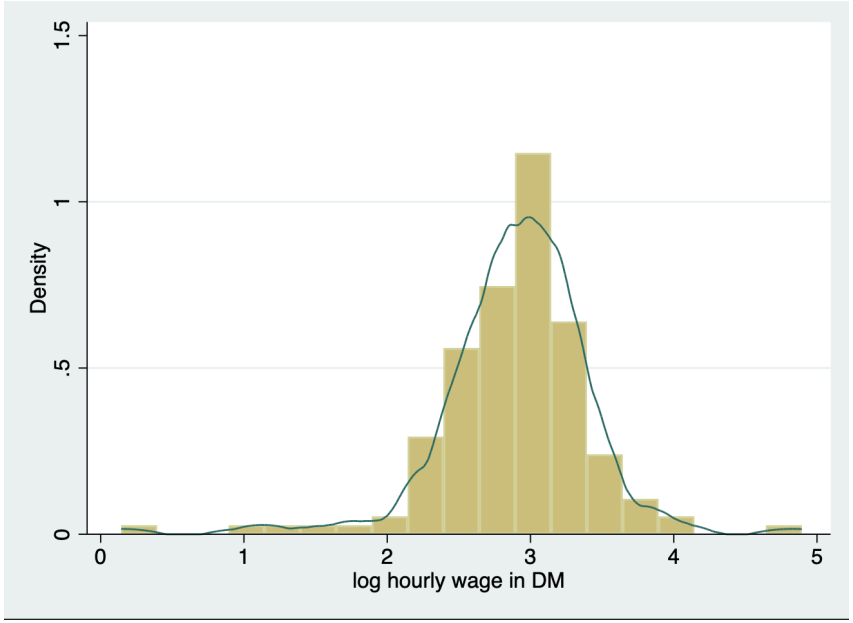
To calculate the density estimate within a bin, all observations inside this bin are given equal weight, and all observations outside the bin are given zero weight.

We are looking for a “better” non-parametric estimate of the density, which is smooth and does not depend on the start and end points of the bins.

Kernel density plot

For data that is continuous or taking many values, a kernel density plot gives a smoother version of the histogram

```
histogram lnw, width(0.25) kdensity
```



- it connects the midpoints of the histogram rather than forming the histogram step function
- it gives more weight to data closest to the point of evaluation

University of Exeter

Kernel density plot

The Kernel density estimator centers a smooth weighting function (Kernel) at each data point of the sample. The contribution of each observed data point to the density at point x is smoothed out over a local neighbourhood.

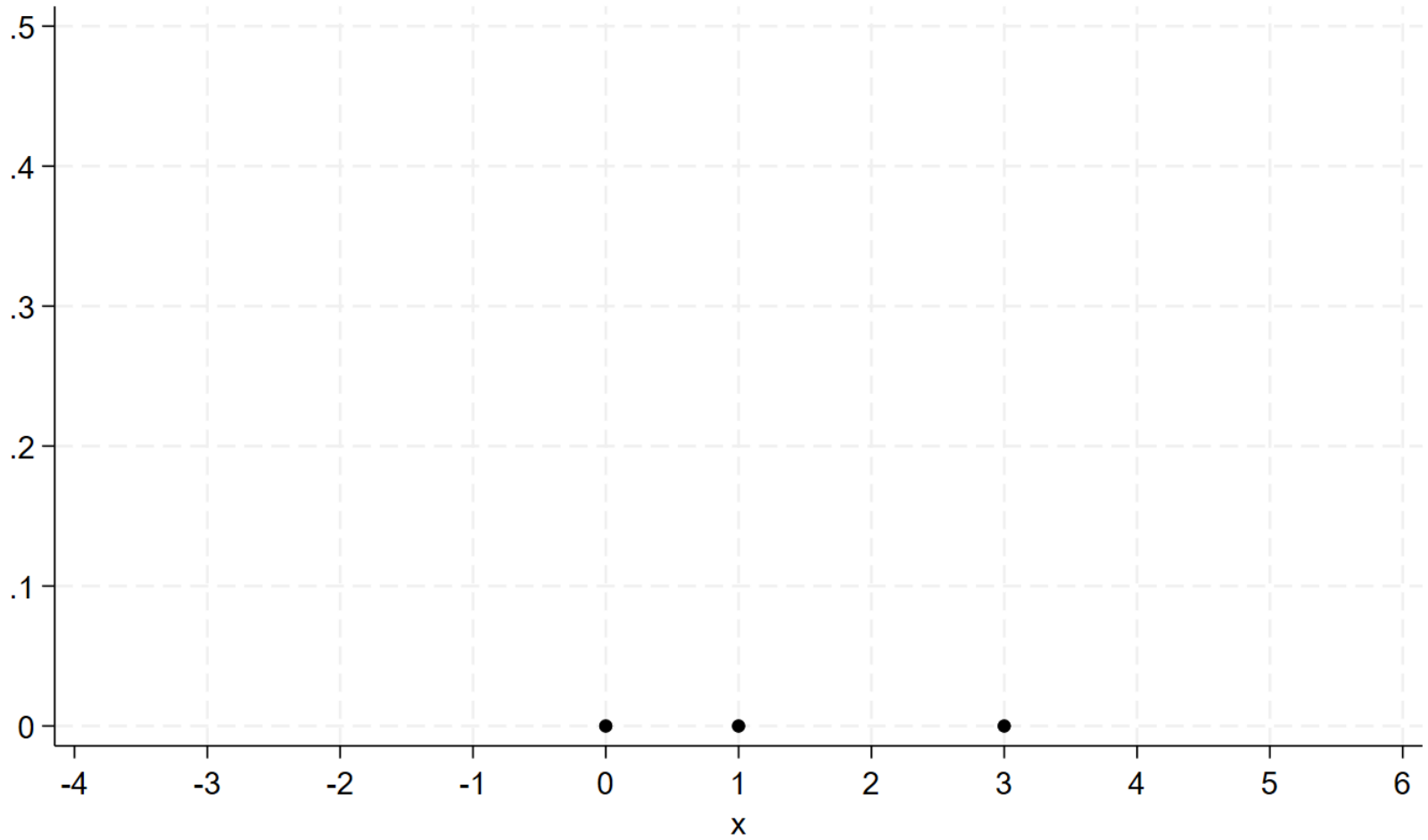
There are 2 parameters to choose:

Kernel A density function that defines the shape of the distribution placed at each point. Default is the Epanechnikov but other can be used.

Bandwidth Defines the "local neighbourhood" (analogue to the width of the bins in a histogram), meaning that it controls the size of the kernel at each point.

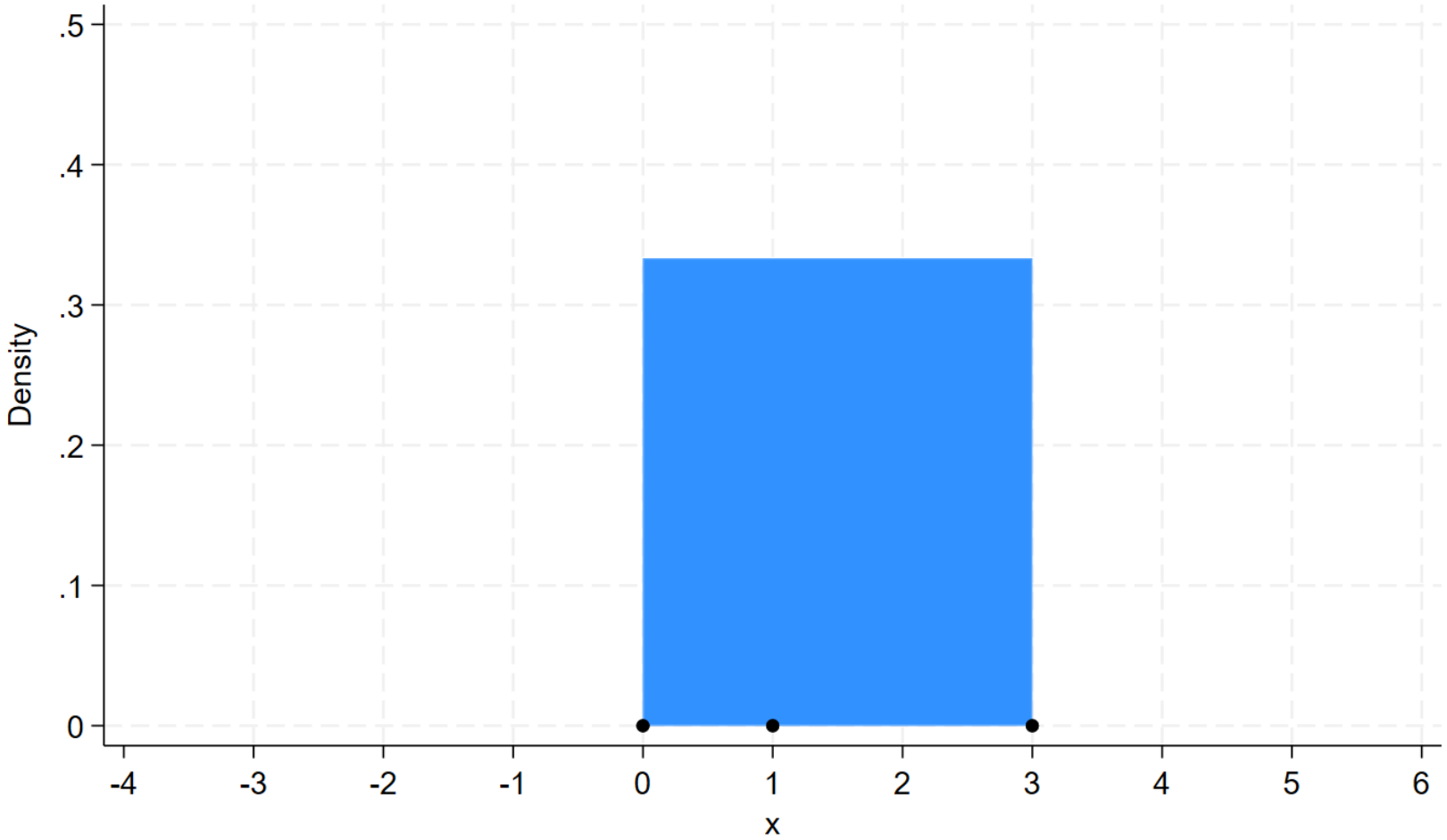
Kernel density plot - example

Suppose there are three data points: X1=0, X2=1, and X3=3.



Kernel density plot - example (cont)

Here is a histogram with three bins.



Kernel density plot - example (cont)

How to construct a kernel density estimate?

Step 1: Parameter choices.

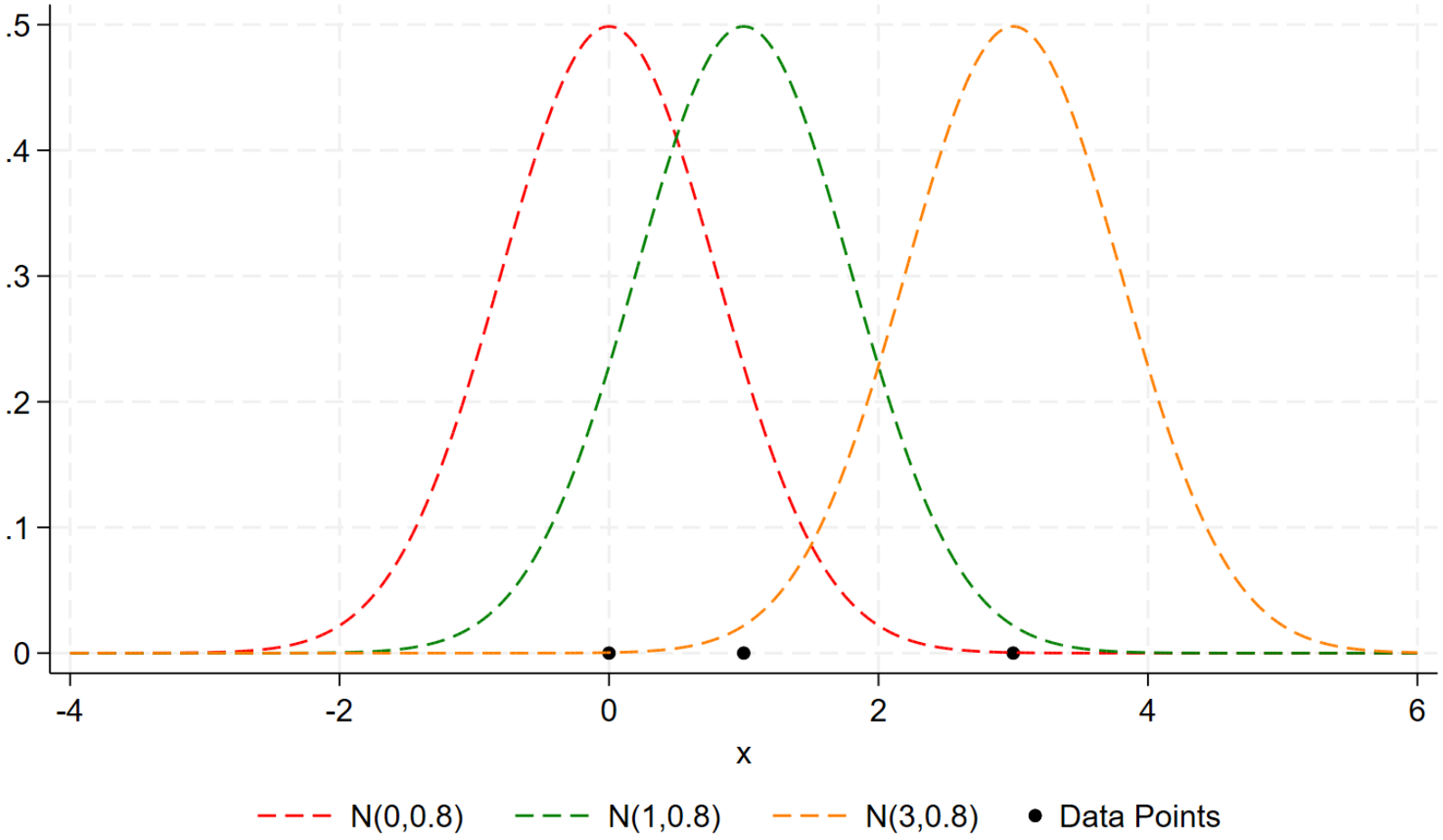
We choose the normal density function for the kernel, and set the bandwidth equal to 0.8.

This translates into a normal density with standard deviation 0.8.

Now we overlay each of our three data points with a N(x,0.8) density function.

Kernel density plot - example (cont)

Step 1:



Kernel density plot - example (cont)

How to construct a kernel density estimate?

Step 2: To find the density estimate for some value of $x = x_0$, we add up all the normal densities from all observations evaluated at $x = x_0$, divided by the number of observations.

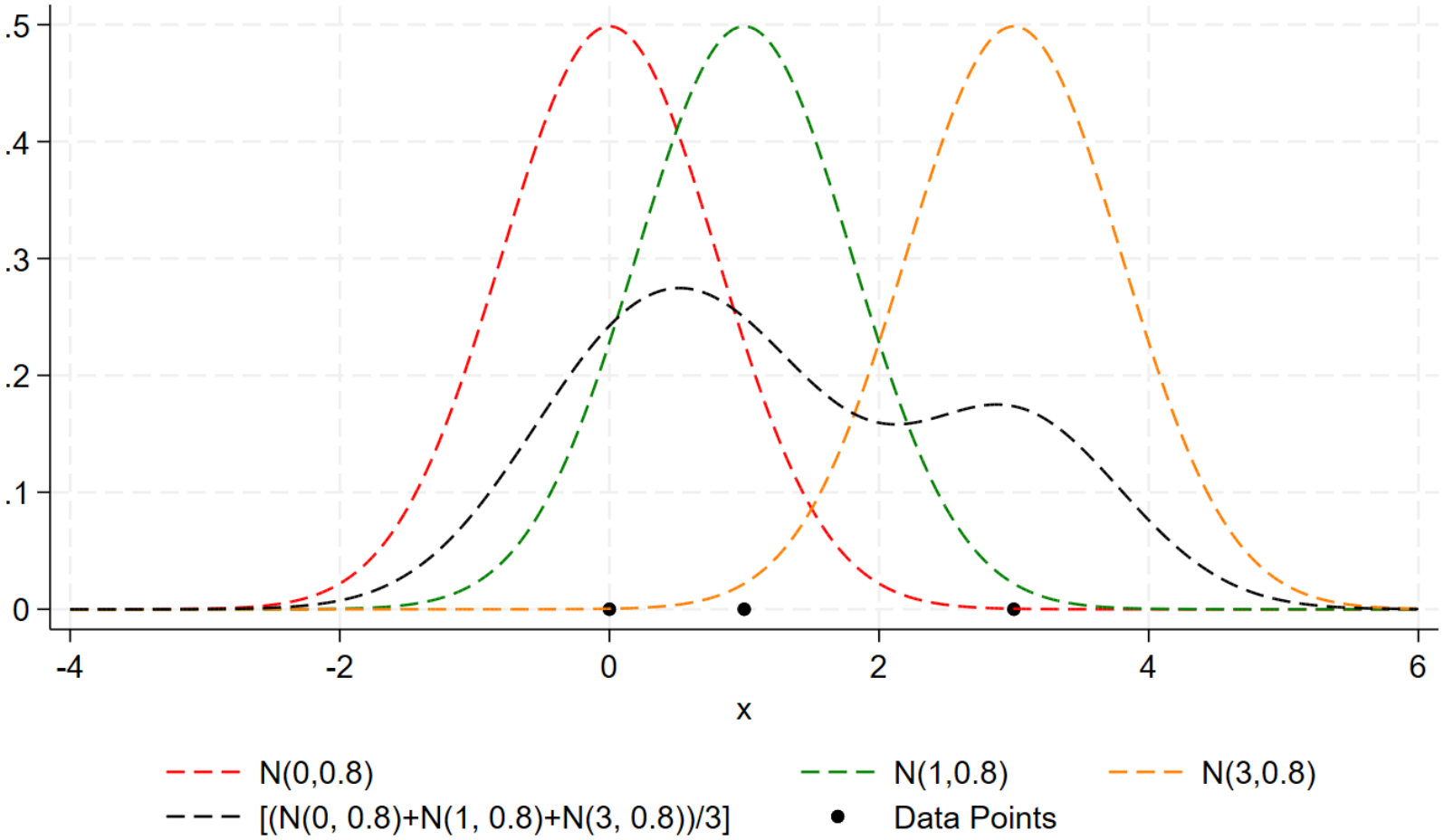
Repeat for as many points x as needed.

Computation example for x1=0:

```
display (normalden(0,0,0.8)+normalden(0,1,0.8)+normalden(0,3,0.8))/3
```

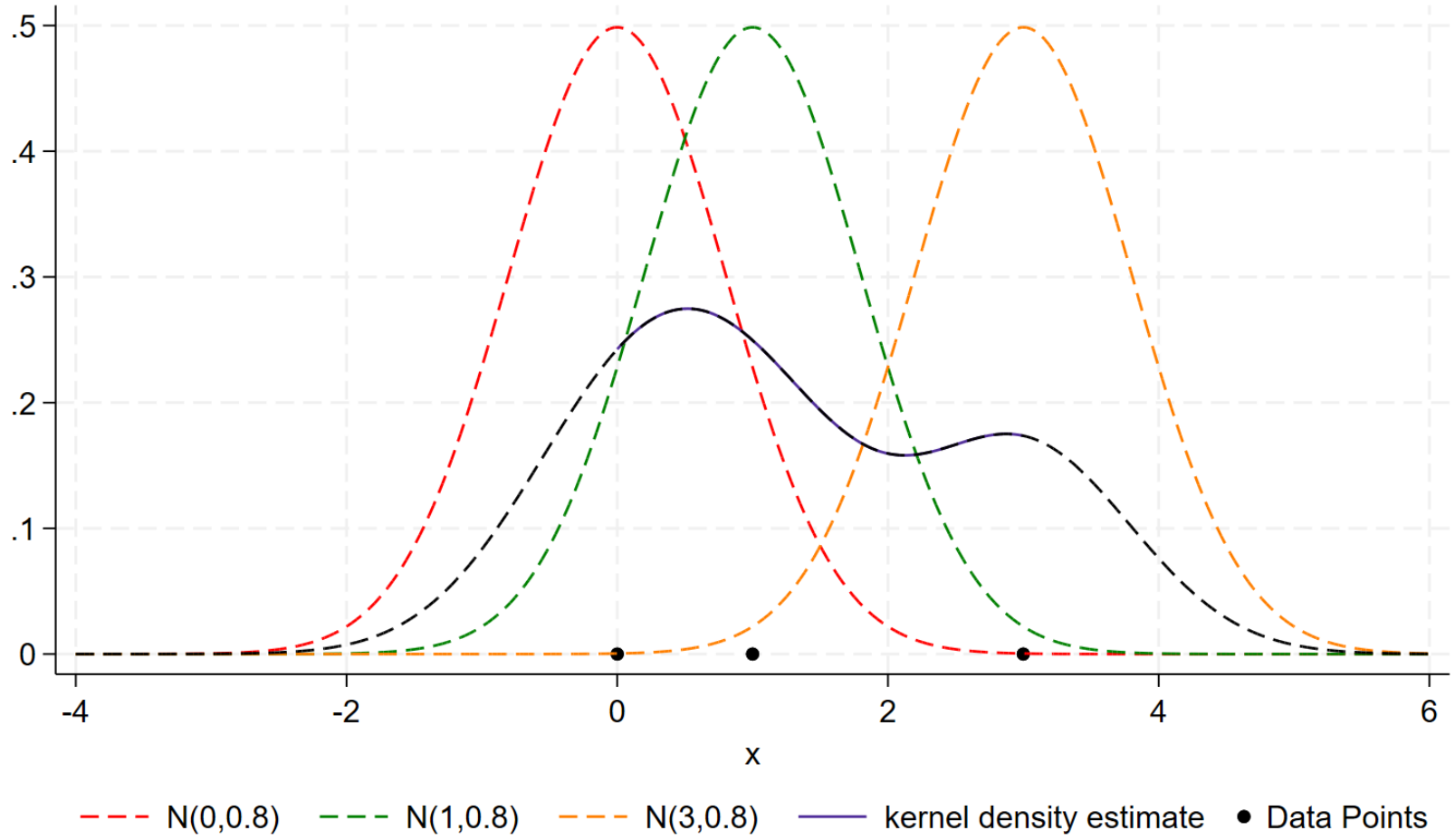
Kernel density plot - example (cont)

Step 2:



Kernel density plot - example (cont)

Check with Stata's `kdensity` command



University of Exeter

Kernel density plot - Choice of Kernel

Good choices of kernels K for continuous data are the Epanechnikov and Gaussian:

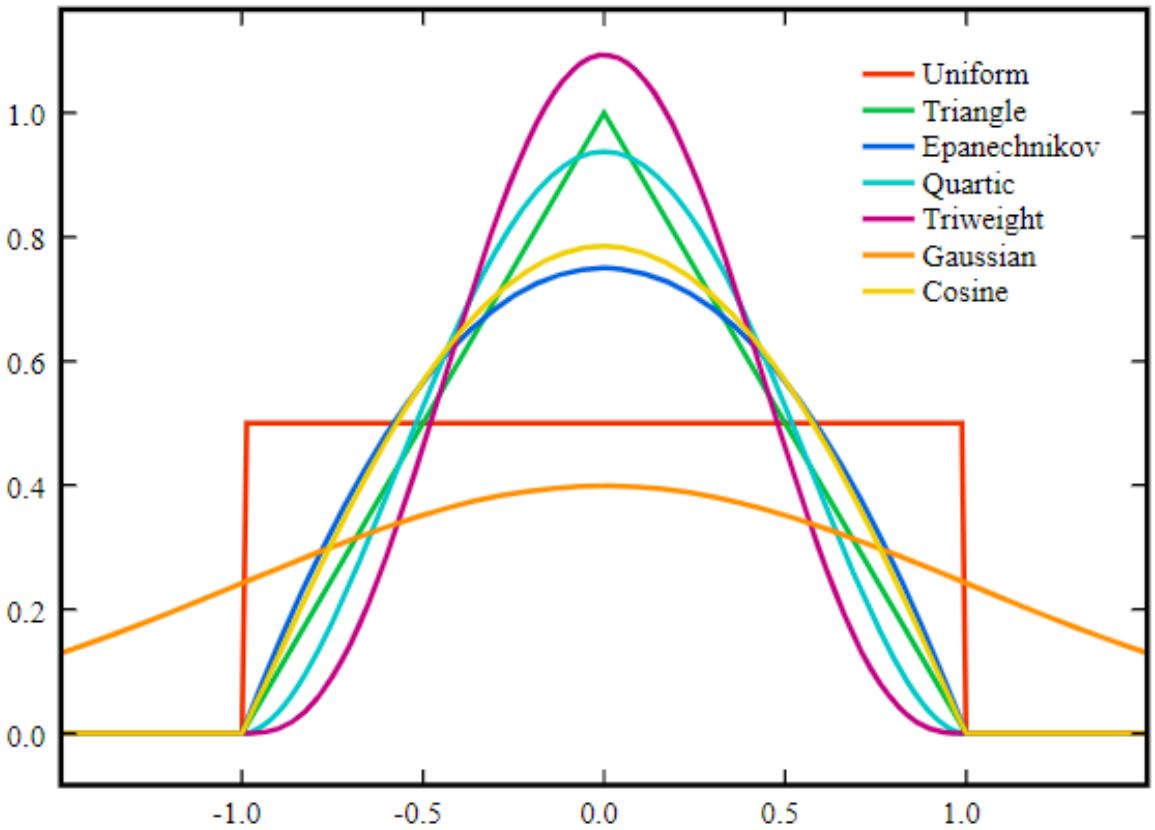
- they give more weight to observations x_i close to the point of evaluation x_0 & provide very similar density estimates
- Epanechnikov is the default kernel in many software because it is AIMSE (asymptotic integrated mean square error) efficient.

Ordering of kernel by smoothness of the representation (if same bandwidth) : 1) cosine is the least smooth, 2) epan2, biweight, triangle, rectangle, parzen, 3) epanechnikov and gaussian are the smoothest.

For instance, Triangular kernel is usually used for nonparametric estimation and not for density estimation as it won't produce a smooth representation.

University of Exeter

Kernel density plot - Choice of Kernel (cont)



Kernel Functions (Source: wikipedia)

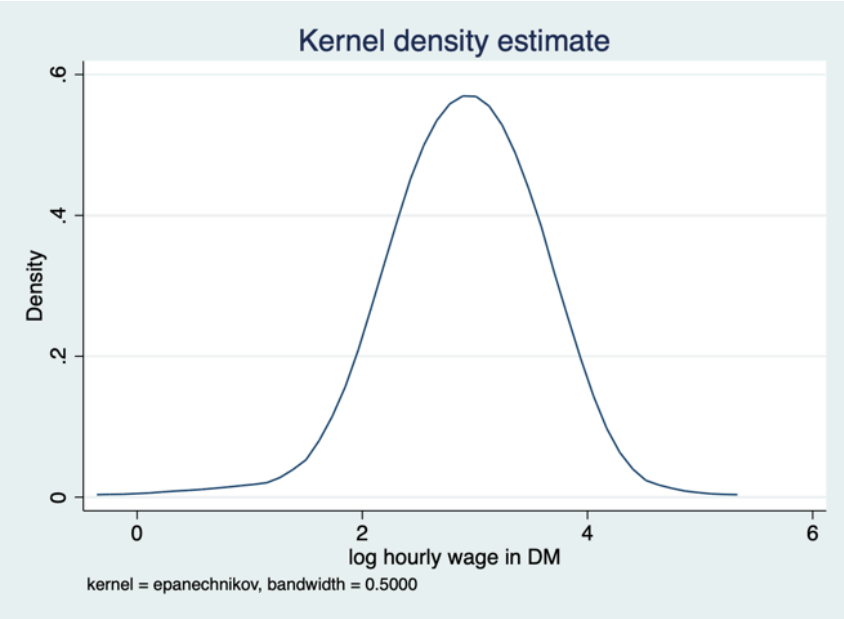
University of Exeter

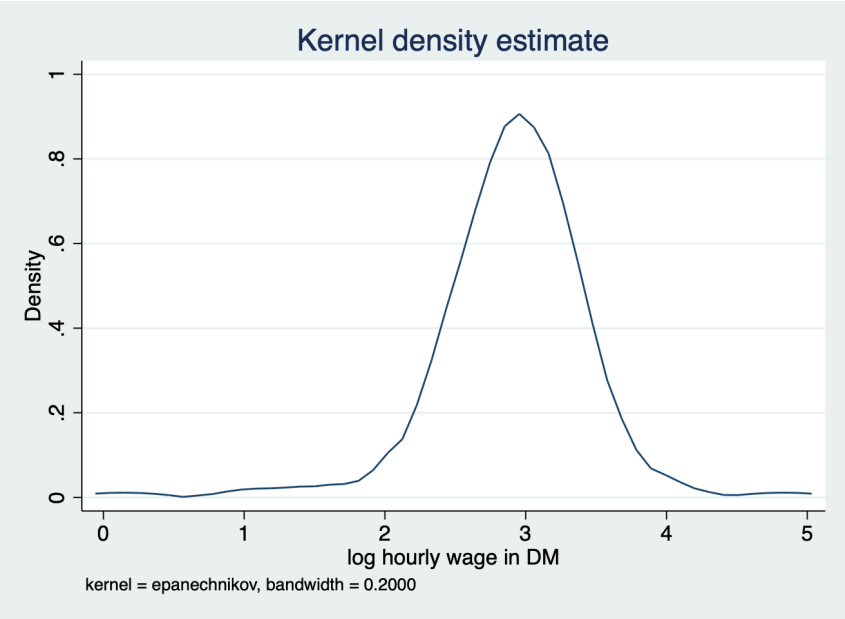
Kernel density plot - Choice of Bandwidth

The bandwidth h is a real number used to control the degree of smoothing: usually, the larger the bandwidth, the smoother the estimators (the representations) are.

See using the `pen01aL3` data

`kdensity lnw, kernel(epanechnikov) bwidth(0.5)`	`kdensity lnw, kernel(epanechnikov) bwidth(0.2)`





University of Exeter

Kernel density plot

Let $\widehat{f}_K(x_0, h)$ denote the kernel density estimate of $\widehat{f}_K(x, h)$ at $x = x_0$

$$\widehat{f}_K(x_0, h) = \frac{1}{Nh} \sum_{i=1}^{N} \mathbf{K} \frac{x_i - x_0}{h}$$

where **K** is a kernel function that places greater weight on points x_i close to x_0, h is the bandwidth, and N the number of observations.

Note that the 3 data points example only serves illustration.

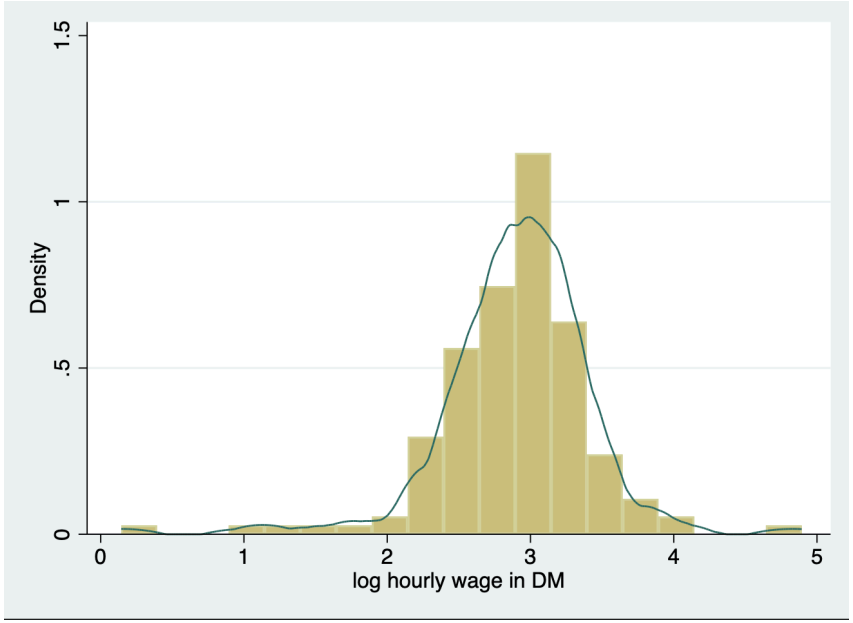
We are left with the choice of h and **K**.

- In large samples, different choices of **K** yield very similar results, whereas the choice of h has a big influence.

Kernel density plot - Comparisons

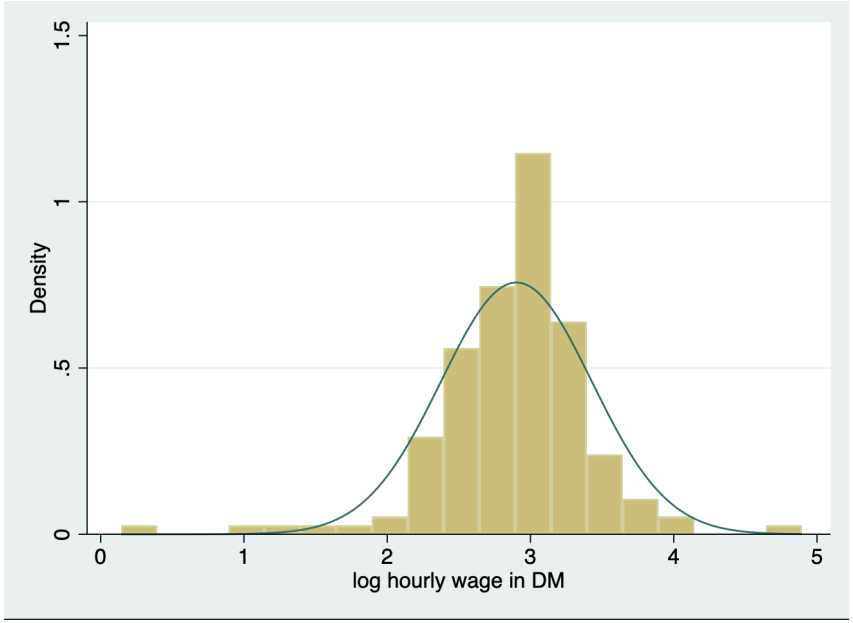
Histogram compared to a kdensity distribution

```
histogram lnw, width(0.25) kdensity
```



Histogram compared to a normal distribution

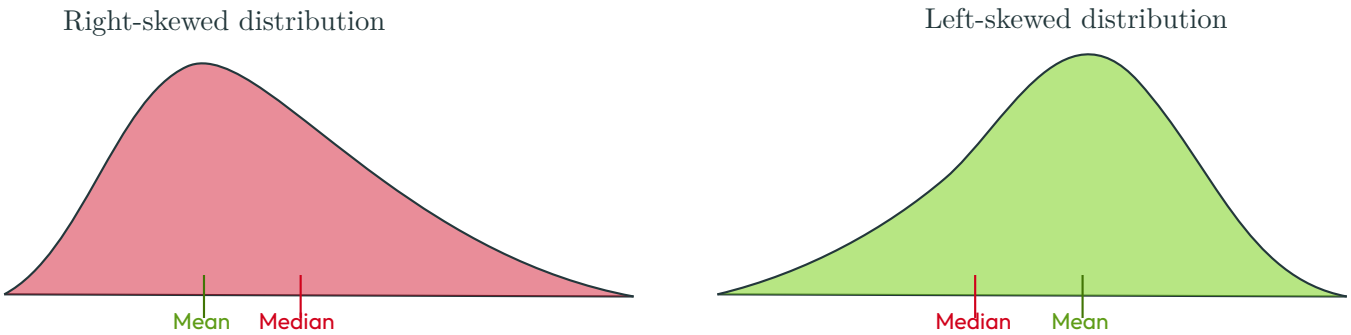
```
histogram lnw, width(0.25) normal
```



University of Exeter

Describing the shape of a distribution

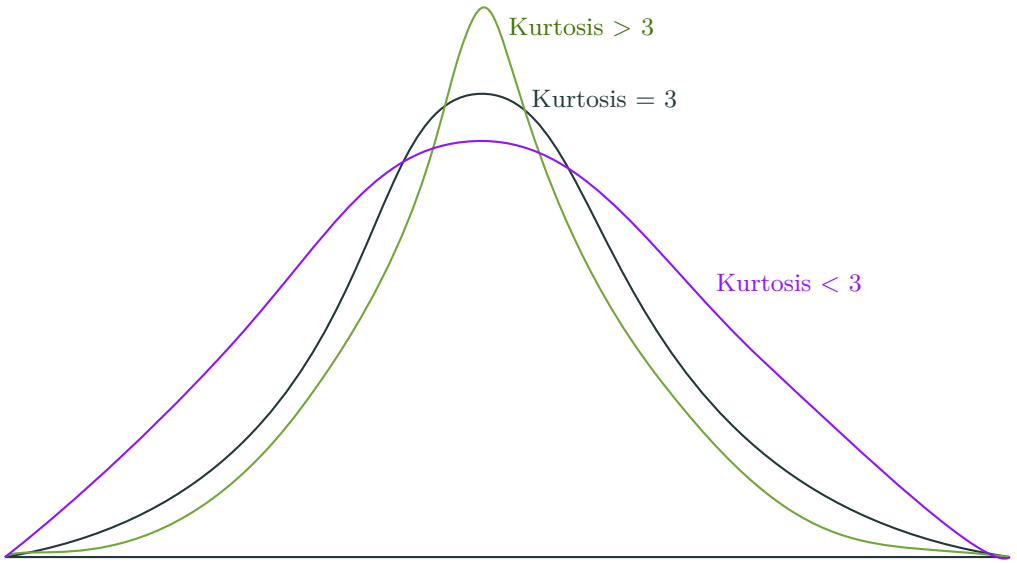
- The shape of the distribution refers to how the observations are distributed around the mean. Are they symmetrically distributed? Are they widely spread around the mean? (are there outliers?)
- The skewness is an indicator of asymmetry:
  - Right-skewed distribution (skewness >0): when outliers pull the mean upward (a few very high values) such that graphically the mean is pulled to the right
  - Left-skewed distribution (skewness <0): outliers pull the mean downwards (a few very low values), such that graphically the mean is pulled to the left.



University of Exeter

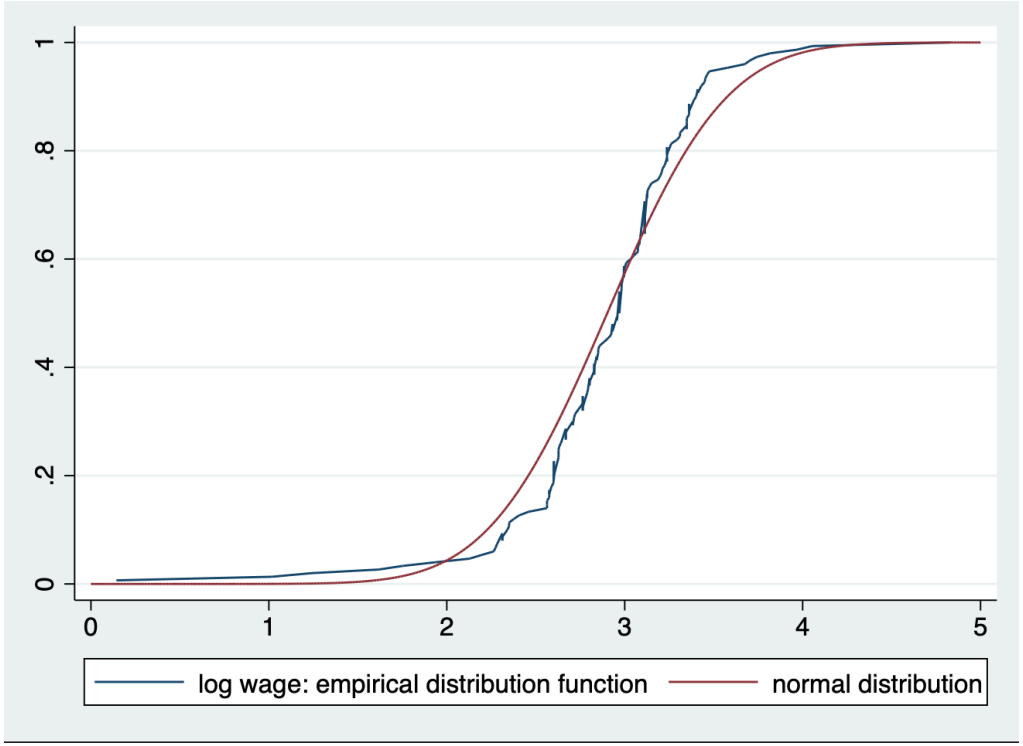
Describing the shape of a distribution

- The kurtosis of the distribution refers to how flat is a distribution, so it indicates the distribution in terms of height of the peak and heaviness of the tails.
  - If kurtosis > 3 : positive kurtosis
  - If kurtosis < 3 : negative kurtosis
  - If kutrosis = 3 : normal distribution



Cumulative distributions

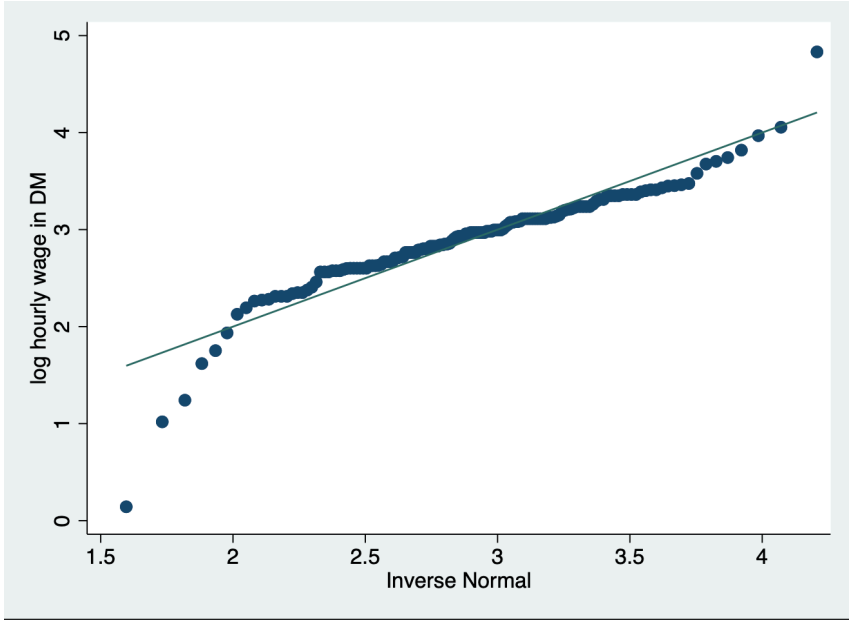
Cumulative distributions provide a graphical assessment of the number/percentage of observations below or equal to a given value or category.



Quantile plots - Quantiles to Normal plot

We can compare the quantiles of a variable with the quantiles of a normal distribution. If the variable follows a normal distribution we should see the point following the line.

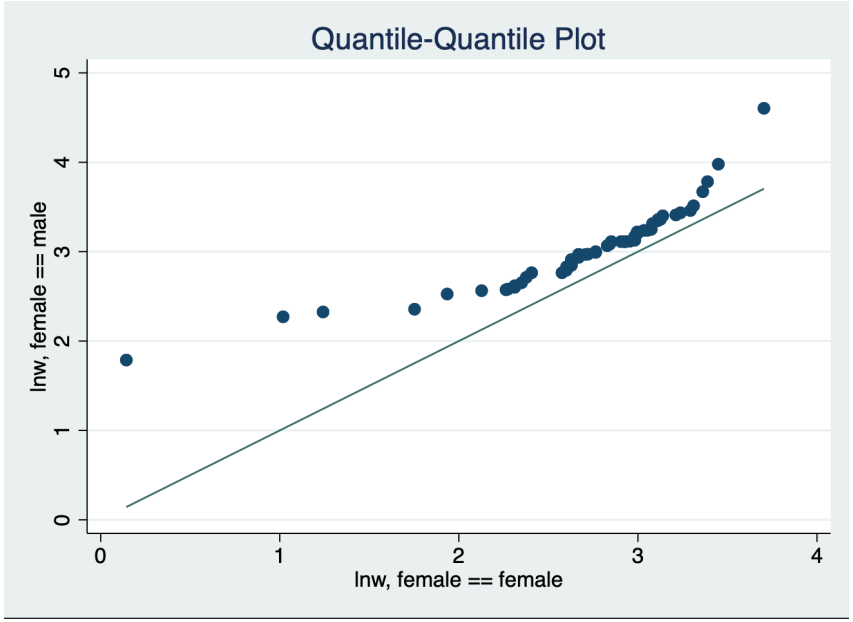
```
qnorm lnw
```



Quantile plots - Quantiles to Quantiles plot

We can also plot the quantiles of one variable against those of another variable. If the variables have the same distribution and follow the line, then we expect a linear relationship between the quantiles.

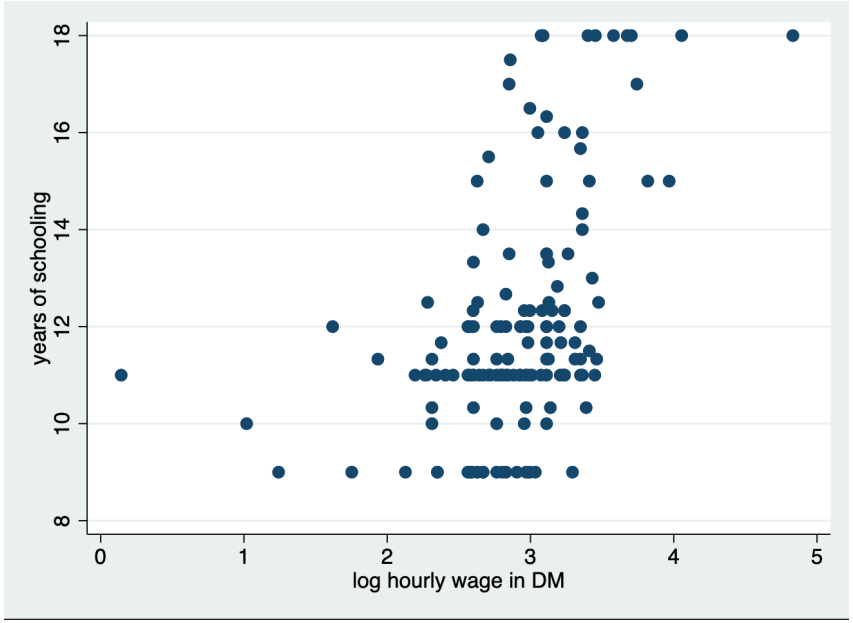
```
seperate lnw, by(female)
qqplot lnw0 lnw1
```



Scatterplots

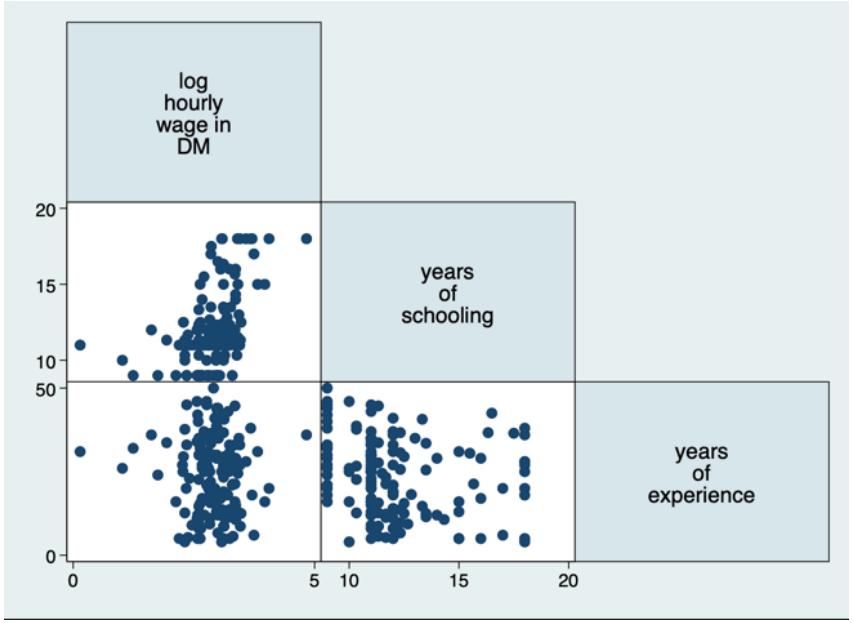
A scatterplot is a convenient way to display all data points in two dimensions:

```
scatter ed lnw
```



We can also combine several scatterplots in a matrix

```
graph matrix lnw ed exp, half
```

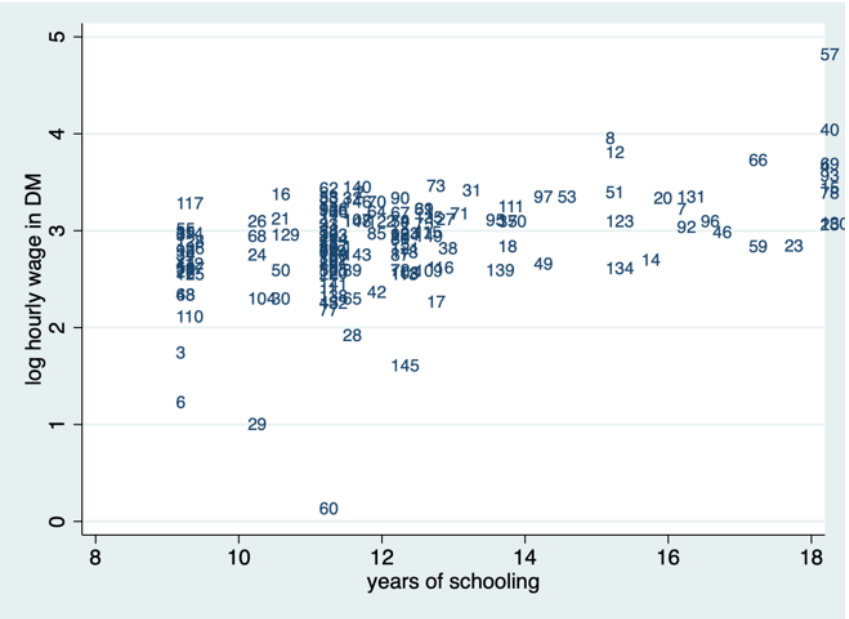


University of Exeter

Scatterplots

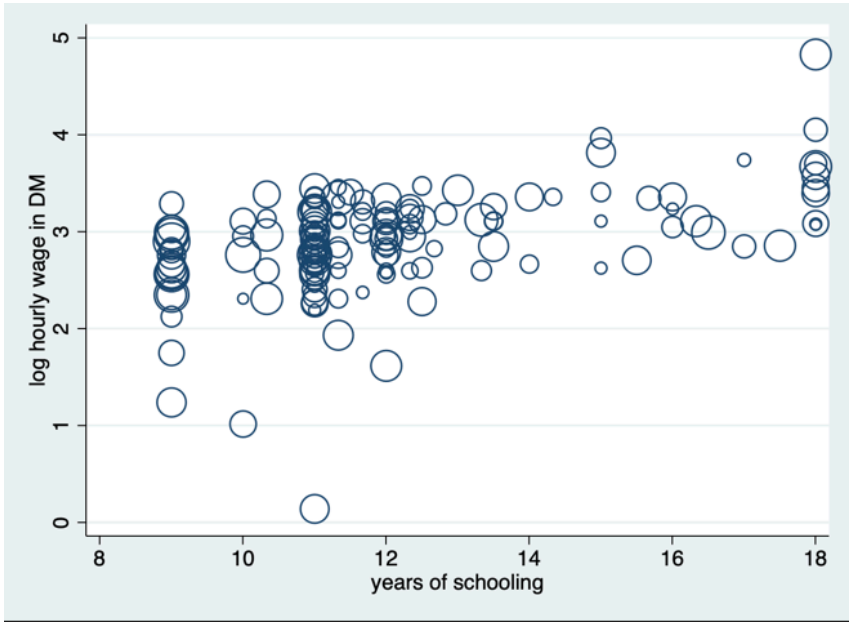
You can include the observation numbers as markers, to help you identify bivariate outliers:

```
gen number = _n
scatter lnw ed, msymbol(i) mlabel(number)
```



You can display the value of a third variable as the size of the markers:

```
scatter lnw ed [aweight=exp], msymbol(oh)
```

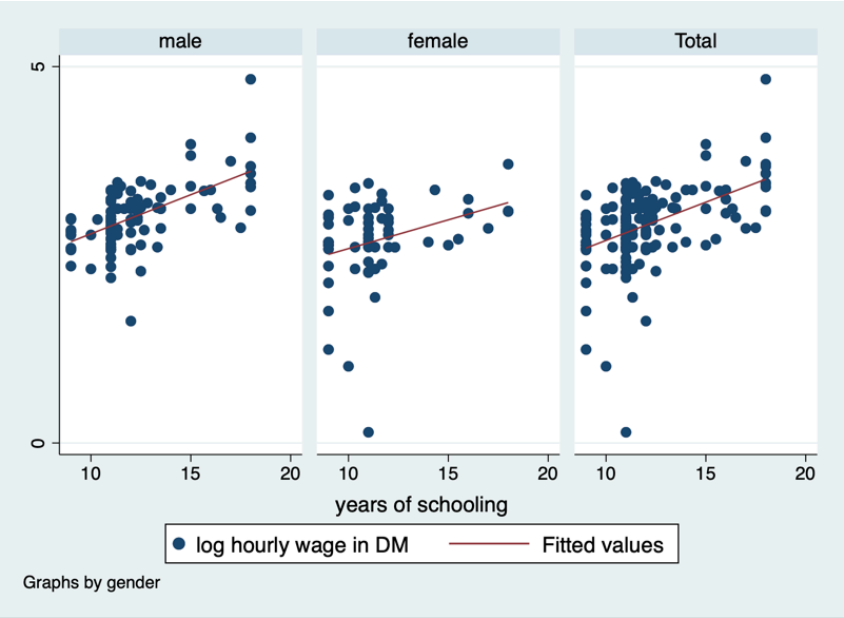


University of Exeter

Scatter plots with linear and quadratic fit

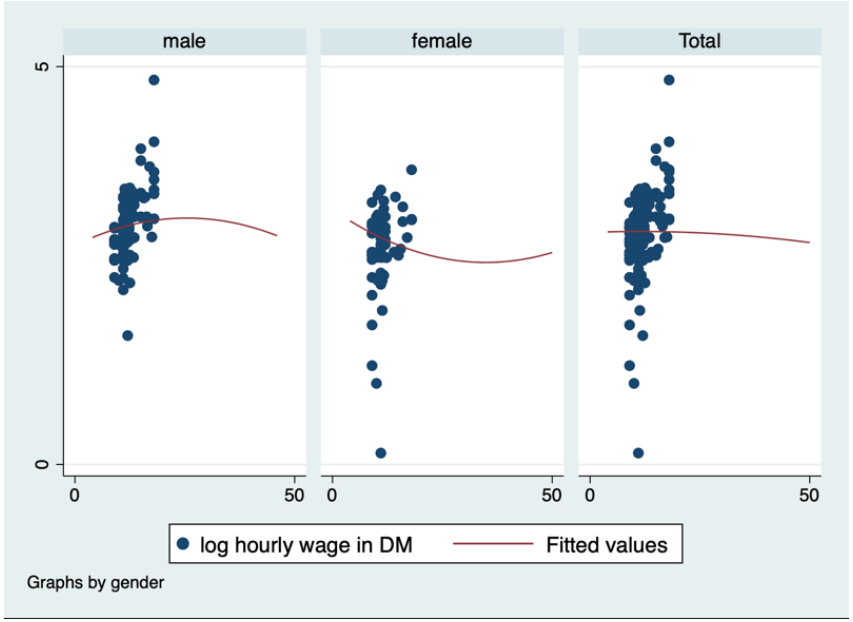
Scatterplot with linear fit:

```
twoway (scatter lnw ed) (lfit lnw ed), ///
by(female, total cols(3))
```



Scatterplot with quadratic fit:

```
twoway (scatter lnw ed) (qfit lnw exp), ///
by(female, total cols(3))
```



University of Exeter

Normality tests

University of Exeter

Normal distribution of the data

If we use a statistical procedure that assumes that the data is normally distributed (many do), we need to test for it.

- **T-tests**, assume that the populations are normally distributed.
- **Linear Regression**, assumes that the residuals are normally distributed.

It's worth noting that with large sample sizes, many parametric tests are robust to violations of the normality assumption due to the Central Limit Theorem.

Normal distribution of the data

The procedure regarding the normality of the data usually follow this order:

- Use graphical methods to investigate normality: histograms/kernels with superimposed normal distribution (see before).
- Use tests to investigate the normality.
- Transform the data and then testing again for normality.

University of Exeter

Testing for Normality

Stata offers 3 buil-in tests for normality:

- Skewness-Kurtosis test, similar to the Jarque-Bera test, for $n \geq 8$: `sktest`
- Shapiro-Wilk test for normality, for $7 \leq n \leq 2,000$: `swilk`
- Shapiro-Francia test for normality, for $5 \leq n \leq 5,000$: `sfrancia`

For these tests, note that the Null hypothesis states that the variable is normally distributed.

Under the Null, the variable has skewness of zero and kurtosis of three.

University of Exeter

Testing for Normality

Using the Skewness-Kurtosis test, which is asymptotically $X^2(2)$ distributed: `sktest lnw`

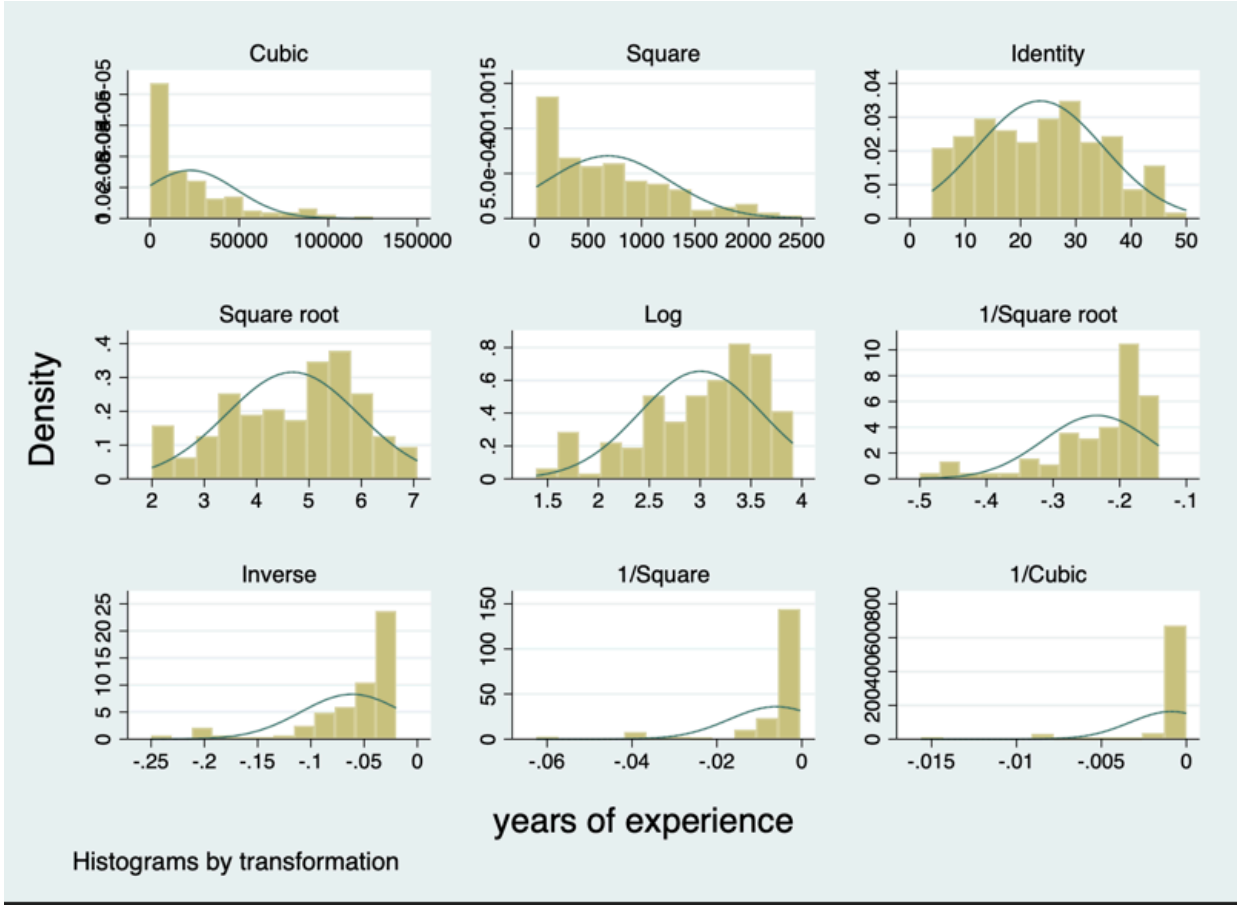
```
. sktest lnw

Skewness and kurtosis tests for normality
                                                    ------ Joint test ------
    Variable |        Obs    Pr(skewness)    Pr(kurtosis)    Adj chi2(2)  Prob>chi2
-------------+-------------------------------------------------------------------
         lnw |        150          0.0000          0.0000          39.93     0.0000
```

High values of the test statistic indicate that the Null hypothesis of normality can be rejected. That is, low p-values (< 0.01, or < 0.005, or < 0.001) provide evidence against normally distributed data. A similar principal apply for the other tests `swilk` and `sfrancia`.

University of Exeter

Transforming the data

We can look at different transformations of the variable to see if there would be transformation that would allow the variable to be closer to a normal distribution: `gladder exp`



Transforming the data

Or we can directly test the normality of transformations: `ladder exp`

```
. ladder exp

Transformation        Formula                chi2(2)  Prob > chi2
------------------------------------------------------------------
Cubic                 exp^3                   35.95        0.000
Square                exp^2                   14.04        0.001
Identity              exp                     13.39        0.001
Square root           sqrt(exp)               10.88        0.004
Log                   log(exp)                12.90        0.002
1/(Square root)       1/sqrt(exp)             31.83        0.000
Inverse               1/exp                   53.75        0.000
1/Square              1/(exp^2)               87.77        0.000
1/Cubic               1/(exp^3)              111.19        0.000
```

Stat Tips Tricks: Saved results

University of Exeter

Working with saved results

Stata not only displays the results but also memorizes it and keep them ready to use after each Stata procedure:

`summarize lnw`

`return list`

```
. summarize lnw

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         lnw |        150    2.901829    .5269156   .1431778   4.831729

.
. return list

scalars:
                  r(N) =  150
              r(sum_w) =  150
               r(mean) =  2.901829149325688
                r(Var) =  .2776400846315046
                 r(sd) =  .5269156333147694
                r(min) =  .1431778073310852
                r(max) =  4.831728935241699
                r(sum) =  435.2743723988533
```

University of Exeter

Saved results: advantages

Precision The saved results are highest possible precision in Stata, whereas the displayed results are usually rounded for ease of display.

No copy/paste Instead of manually typing numbers into Stata, we can use saved results to avoid mistakes.

Efficient coding Using saved results (assigning it a name) allow to re-use our commands using different variables.

Note that returned results are usually only active until another Stata command that also returns results is issued.

University of Exeter

Saved results: example

In Stata commands, scalars are referred to simply by their name.

Example

- To display the squared standard deviation: `display r(sd)^2`
- To remove the mean: `generate demean = lnw - r(mean)`
- `generate standard = (lnw - r(mean))/r(sd)`

Application: Jarque-Bera test for normality

$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4} X^2(2)$$

where S is the skewness, K the kurtosis, and n the number of observations.

University of Exeter

Saved results: example

Hard way:

summarize lnw, detail

display (150/6)*((-1.150735)^2 + (9.284479 − 3)^2/4)

```
. summarize lnw, detail

                    log hourly wage in DM
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.018646       .1431778
 5%     2.194448       1.018646
10%     2.345427        1.24179       Obs                 150
25%     2.631166       1.619084       Sum of wgt.         150

50%     2.969492                      Mean           2.901829
                        Largest       Std. dev.      .5269156
75%     3.199535       3.818812
90%     3.406059       3.968189       Variance       .2776401
95%     3.579777       4.055201       Skewness      -1.150735
99%     4.055201       4.831729       Kurtosis       9.284479

. display (150/6)*((-1.150735)^2 + (9.284479-3)^2/4)
279.9465
```

Saved results: example

Better way:

```
summarize lnw, detail
return list
display (r(N)/6)*((r(skewness))^2 + (r(kurtosis) - 3)^2/4)
```

```
. summarize lnw, detail

                    log hourly wage in DM
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.018646       .1431778
 5%     2.194448       1.018646
10%     2.345427        1.24179       Obs                 150
25%     2.631166       1.619084       Sum of wgt.         150

50%     2.969492                      Mean           2.901829
                       Largest       Std. dev.      .5269156
75%     3.199535       3.818812
90%     3.406059       3.968189       Variance       .2776401
95%     3.579777       4.055201       Skewness      -1.150735
99%     4.055201       4.831729       Kurtosis       9.284479

. return list

scalars:
                  r(N) =  150
              r(sum_w) =  150
               r(mean) =  2.901829149325688
                r(Var) =  .2776400846315046
                 r(sd) =  .5269156333147694
           r(skewness) =  -1.150735466513915
           r(kurtosis) =  9.284479436392603
                r(sum) =  435.2743723988533
                r(min) =  .1431778073310852
                r(max) =  4.831728935241699
                 r(p1) =  1.018646001815796
                 r(p5) =  2.194447994232178
                r(p10) =  2.345427393913269
                r(p25) =  2.631165981292725
                r(p50) =  2.969491958618164
                r(p75) =  3.199534893035889
                r(p90) =  3.40605902671814
                r(p95) =  3.579777002334595
                r(p99) =  4.055201053619385

. display (r(N)/6)*((r(skewness))^2 + (r(kurtosis)-3)^2/4)
279.94656
```

Saved results: additional elements

- Remember to use saved results directly after the corresponding command. Later commands will erase them.
- Local macros and scalars are convenient way to store the results; we will learn about them later.
- You need to use `return list` only if you are unsure whether a specific result is returned, or how it is named. Once you have memorized `r(N)`, `r(mean)` tec. there is no need to call `return list`; simply use them in your commands.
- Trye `return list` on other commands as well, e.g. after `correlate`, `kdensity`, `sktest`, `ladder`.