STAT 518 --- Nonparametric Density Estimation

• The <u>probability density function</u> (or <u>density</u>) of a continuous random variable X describes its probability distribution.
• We denote the density as
• Note that if F(x) is the c.d.f. of X, then

Two important properties of density functions

(1) They are always _________________:

(2) The total area under a density curve is always _____.

• In real data analysis, we do not know the true density, so we can estimate it using sample data $X_1, X_2, \ldots, X_n$.

<u>Parametric approach</u>: Assume a specific functional form (e.g., normal, gamma, etc.) for the density and use the sample data to estimate certain _________________.

<u>Example</u>: Could assume the density is <u>normal</u> and get sample estimates of _____ and _______.

• The <u>nonparametric</u> approach is to make very <u>few</u> assumptions about the functional form of the density.

Histograms

• A simple density estimator is a <u>histogram</u>.

• In introductory statistics, we study the _____________ histogram having bins with bars whose height is the count of sample observations falling in that bin.

• If we rescale the heights of each bar so that the <u>total combined area</u> within all the bars is 1, we have a <u>histogram density estimate</u>.

• Assume there are K bins, each of width h:

Picture (K = 5, h = 2):

• In general, this histogram is:

where

• **The total combined area within all bars is**

• **The R function `hist` produces such histograms.**

• **The choice of bin width *h* determines the number of bins, which can affect the appearance of the estimate.**

• **A simple rule of thumb for choosing *h* is derived from a normal density:**

Let

where

• **Note: the sample standard deviation *s* is a consistent estimator of σ, as is *IQR* / 1.34 when the true density is normal.**

• **In reality, this provides a good initial choice of *h*, which may then be adjusted by trial and error.**

• **Choosing *h* too small produces <u>many bins</u> and a density estimate that is too ____________.**

• **Choosing *h* too large produces <u>few bins</u> and a density estimate that is ____________________.**

Example 1:

Example 2:

• **We could also let the bin width vary across bins, choosing a ________ width in regions where we expect the density to be <u>flatter</u> and a ________ width in regions where we expect the density to be <u>spiky</u>.**

Kernel Density Estimation

• **An obvious drawback to the histogram density estimate is that it is not ____________.**

• **A <u>kernel density estimate</u> (k.d.e.) produces a smooth estimate and works similarly to the kernel regression method.**

• **As $n \to \infty$, the k.d.e. will approach the true density f(x) more quickly than the histogram will.**

Recall:

• Plug in the e.d.f. for $F(\cdot)$ to obtain:

• This is exactly the same as

with $K(u)$ =

→ a kernel estimate with a __________ kernel function.

• However, with the ____________ kernel, the resulting density estimate is not smooth.

• Better choices of kernel function $K(\cdot)$ include:

• Let $K(\cdot)$ in the above k.d.e. formula be a standard normal kernel function.
• Then for, say, h = 1:

• We see at each point *x*, the k.d.e. is the average of normal densities, centered at

• Sample values near *x* will contribute

• Sample values far from *x* will

Role of the Bandwidth *h*

• If *h* increases, these normal densities become ________ and more ___________________

→

• If *h* decreases, these normal densities become _______ and ___________________

→

• Rule of thumb for choosing *h* (again based on the true density being normal):

Let

where

• In reality, this provides a good initial choice of *h*, which may then be adjusted by trial and error.

• The `density` function in R produces a kernel density estimate.

Example 1:

Example 2:

• As with kernel regression, kernel density estimators tend to be biased at the left and right edges:

• The k.d.e. also has a tendency to be too flat (not rise or dip enough) in the peaks and valleys of the density.

• An option is to use a bandwidth that <u>varies</u> over the region (being ___________ where the density is expected to be flat and ____________ where the density is expected to have bumps).