Week 1. Central tendency, Q-Q plot, lognormal distribution, and kernel density estimation

Three central tendency measures and their domain of application. Q-q plot for testing the distribution assumption and the confidence band. The hymn to the lognormal distribution and the CLT on the log/relative scale. Kernel density estimation (`density`) for the analysis for hourly wages in U.S. Application of kernel density to toenail arsenic distribution and estimation of the hazard health threshold in NH.

R codes: `qqShape, qqband, toears, jackM, kern.movie`

Data: `toears.txt, height1000.csv`

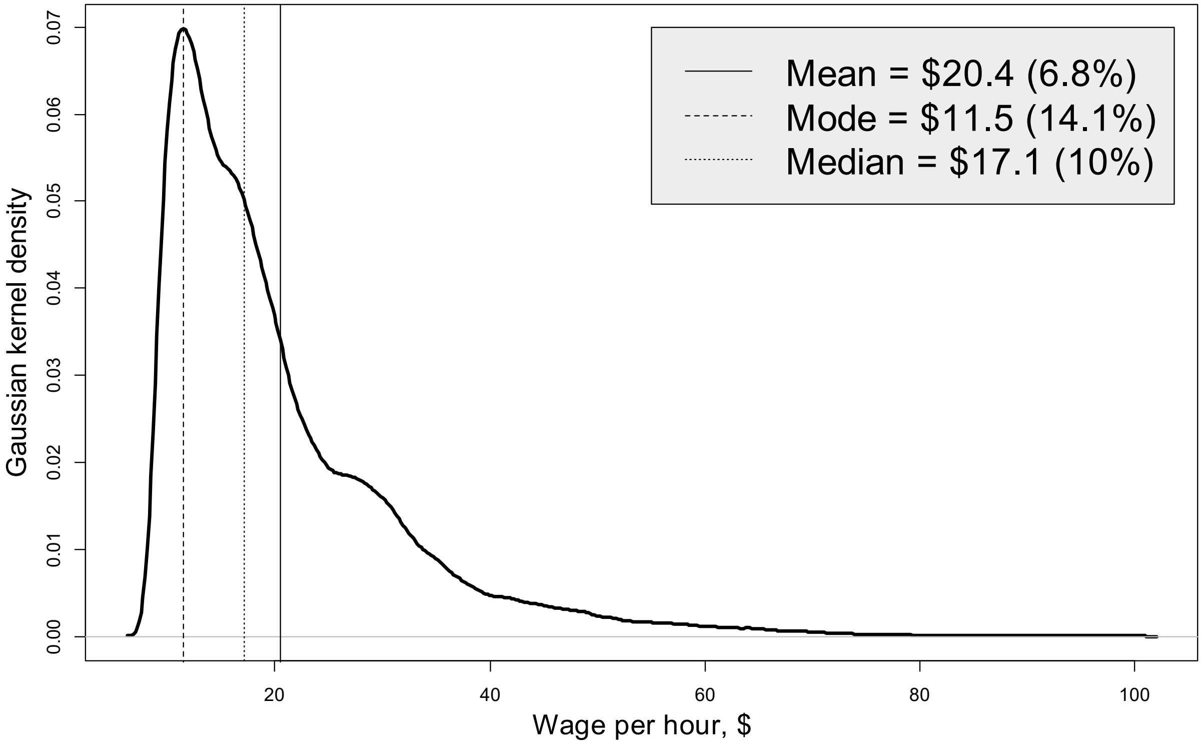
Overview of three central tendency measures

Section 1.4

My goal is to plant a seed of doubt in your mind on the wide-spread use of the mean to report the center of the distribution. Reporting a central tendency measure, such as mean, mode, or median, is the most frequent task of data science.

Median is better but it reflects the comparison at the middle point of the data.

Mode is the point of maximum concentration of the data, that is, where the density reaches its maximum



The Gaussian kernel density for a sample of 234,986 hourly wages in the country.

Example 1 *What central tendency measure to use for a house price in town? Mean, median, or mode?*

Case 1. Mean. The arithmetic average of house prices is a suitable average characteristic for a town clerk who is concerned with the total amount of the property tax to collect from the residents.

Answer. Arithmetic average = Total value of houses in town/number of houses. When reporting the average house price, town officials prefer to use the arithmetic average because they collect the property tax proportional to Total value of houses in town.

Case 2. Median. A real estate agent shows houses to a potential buyer. What is a suitable measure for house price for the buyer, the mean or the median?

Answer. While the mean price makes sense for a town or state official but it is not useful for the buyer who is thinking of the chance of affording the house he/she likes. Instead, the median means that 50% of the houses he/she saw will have price lower than the median and 50% of the houses will have higher price. In this case, the median has a better interpretation from the buyer's perspective.

Case 3. Mode. What central tendency to use to reflect the distribution of house prices in town on the real estate webpage?

Answer. "The most typical house price sold in town is from $350K to $400K." The mode.

Conclusion 2 *Mean has an interpretation if and only sum has an interpretation.*

Example 3 *A hand watch making company wants to know the most fittable wrist length to produce. Having a number of wrist measurements what measure of central tendency, mean, mode, or median, would you use? Justify the answer.*

Median can be computed in R in two ways:

- Using a built-in function `median`
- Sorting observation in ascending order using `sort` command and then extracting the $n/2$ th element: `soX=sort(X);medX=soX[length(X)/2]`

There is no function in R to compute the mode!

If X is the data vector compute the Gaussian kernel density as `d=density(X)`. It returns a data frame with columns `d$x` and `d$y`. Then compute the mode as `d$x[d$y==max(d$y)]`, that is the value x where the density estimate is max.

Q-Q plot

Section 5.3

How to test that your data came from a theoretical distribution? Plot two cdfs.

Section 2.1.

Definition 4 *CDF (cumulative distribution function), F, is the probability that a continuous random variable X takes value equal to x or smaller*

$$F(x)=\Pr(X\leq x)=\int_{-\infty}^{x}f(t)dt,\quad -\infty<x<\infty$$

where f is the pdf (probability density function). For discrete RV

$$F(x)=\frac{1}{n}\sum_{i=1}^{n}1(X\leq x_i)=\frac{1}{n}\#(X\leq x),\quad -\infty<x<\infty$$

where X takes n distinct values $x_1,...,x_n$.

For continuous RV cdf we have (a) $F(-\infty)=0$ and $F(\infty)=1$ and (b) $F'=f$ and F is an increasing function.

Definition 5 *Quantile is inverse cdf. The pth quantile, $q=q_p$ is the solution of the equation*

$$F(q)=p.$$

The pth quantile of a discrete distribution (data array) is the pth order statistic $X_{(np)}$.

Assumption on the **normal distribution** is commonplace. How to test that the data came from a normal distribution?

Question: how to prove that two distributions are the same? Limitations of the mean and the t-test.

Plot quantiles of the standard normal distribution (x-axis) versus empirical quantiles (ordered observations). If the data are normally distributed you must see a straight line.

How to interpret a q-q plot? See R function `qqShape`

How to plot q-q confidence band?

See R function `qqband`

The hymn to the lognormal distribution

Sections 2.10.2, 2.11, 5.4

Things around us bear positive values. This contradicts normal distribution.

Central Limit Theorem on the log scale.

In many situations, contributing factors, X_i, are not **additive**, but **multiplicative**. This phenomenon gives rise to the **lognormal** distribution (discussed further in the next section). In this section, we explain why many real-life quantities follow this distribution due to the multiplicative effect.

The traditional CLT assumes additivity: the difference between two consecutive sums, $S_{i+1} = \sum_{j=1}^{i+1} X_j$ and $S_i = \sum_{j=1}^{i} X_j$, does not depend on S_i because

$$S_{i+1} - S_i = X_{i+1}.$$

We say that CLT relies on additivity. One of the consequences of the additivity is the symmetry of the normal distribution. Consequently, the normal distribution may take negative values with positive probability. In fact, violations of additivity are seen in almost every distribution we deal with in real life; below are just a few examples. In these examples, and in many others, contributing factors act on the **relative** (or equivalently, multiplicative) scale.

1. *The weight of humans, or other biological growth.* Of course genetics play an important role, but a fat-filled diet and unhealthy lifestyle contribute to obesity on the relative scale $Q_{i+1} = Q_i(1 + X_i)$. Obese people gain more weight compared with people with low weight, where X_i is positive or negative but relatively small. The same holds for any other biological growth: factors apply to the current state and the change is proportional to Q_i.

2. *Salary and income.* Salary rises on the relative (or percent scale), $Q_{i+1} = Q_i(1 + X_i/100)$, where X_i is relatively small and represents the percent salary increase. The same rule applies to wealth.

The lognormal distribution is the distribution of

$$\exp \mathcal{N}(\mu, \sigma^2).$$

Many real-life data become normally distributed after the log transformation

Example 6 ***Arsenic toenail distribution in New Hampshire.*** *Arsenic belongs to a group of toxic metals and may cause cancer. Several epidemiology papers relate the elevated concentration of arsenic in drinking water and the resulting excessive toenail arsenic concentration to bladder cancer. The distribution of arsenic in toenails may help environmental policymakers determine the threshold of the level of concentration above which the exposure to arsenic becomes dangerous for health. Use histogram and kernel density for the original and log10 data on toenail arsenic distribution among 1,057 healthy New Hampshire residents. Identify the health-hazard toenail arsenic threshold.*

Solution. See R function `toears`

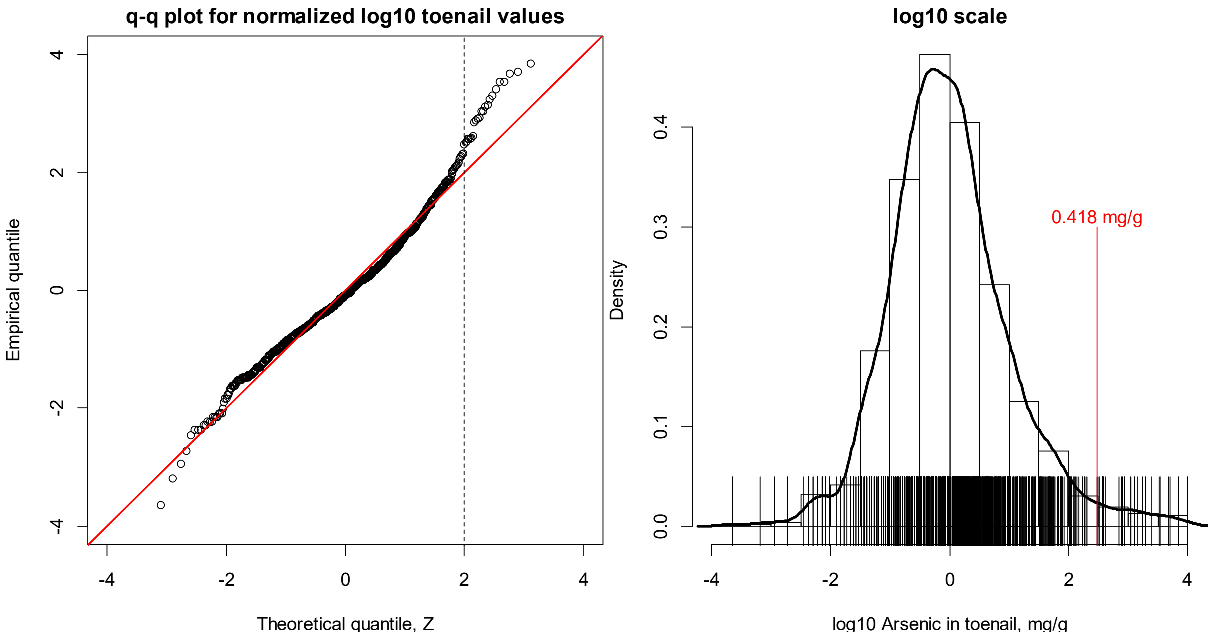


Figure 1: *The q-q plot to test the normality and the histogram with a larger number of bins. The hazard heath threshold is 0.418 mg/g.*

Kernel density estimation

Section 5.5

Let $\{X_i, i = 1, 2, ..., n\}$ be a random (iid) sample from a population with unknown density. Let h be a fixed positive parameter, called the *bandwidth*. We estimate the density as the average of n local kernel densities with center at X_i and SD $= h$, or more specifically as

$$f(x; h) = \frac{1}{nh} \sum_{i=1}^{n} \phi \left(\frac{x - X_i}{h} \right), \tag{1}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

is called kernel (zero mean and unit variance). The difference between histogram and density estimation: `breaks` versus bandwidth (`bw`).

See R function `density`
Run `kern.movie` to see how bandwidth(bw) affect the density estimation for four kernels:

1. The larger the bw the smoother the density.

2. The choice of the kernel is not important, the normal kernel is a good one.

The R function `density` returns the list with names `$x` and `$y`

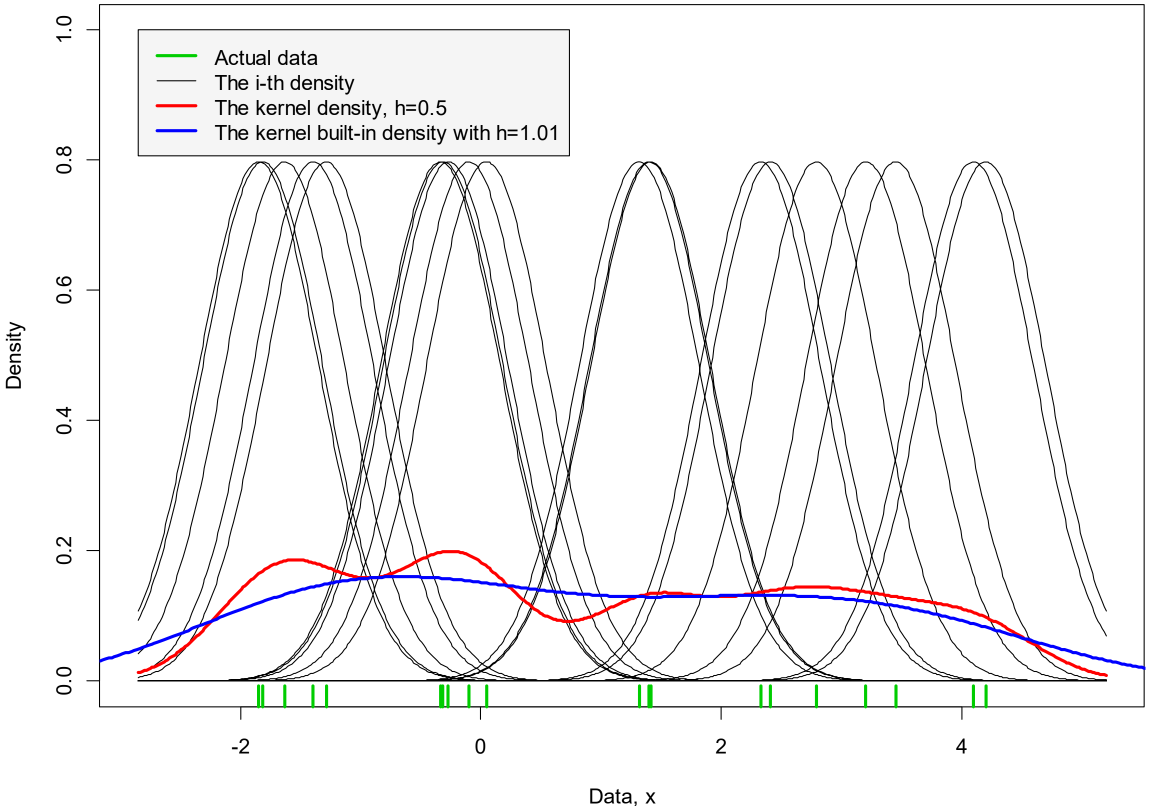


Figure 2: *Kernel density estimation is the average on n local densities. The ith local density has mean X_i and the same SD $= h$. The actual data are shown as bars on the x-axis (`rug` command).*

Mode estimation

```
pdf_X=density(X)
mode_X=pdf_X$x[pdf_X$y==max(pdf_X$y)]
```

The following example illustrates how to use kernel density for estimation of kernel cdf.

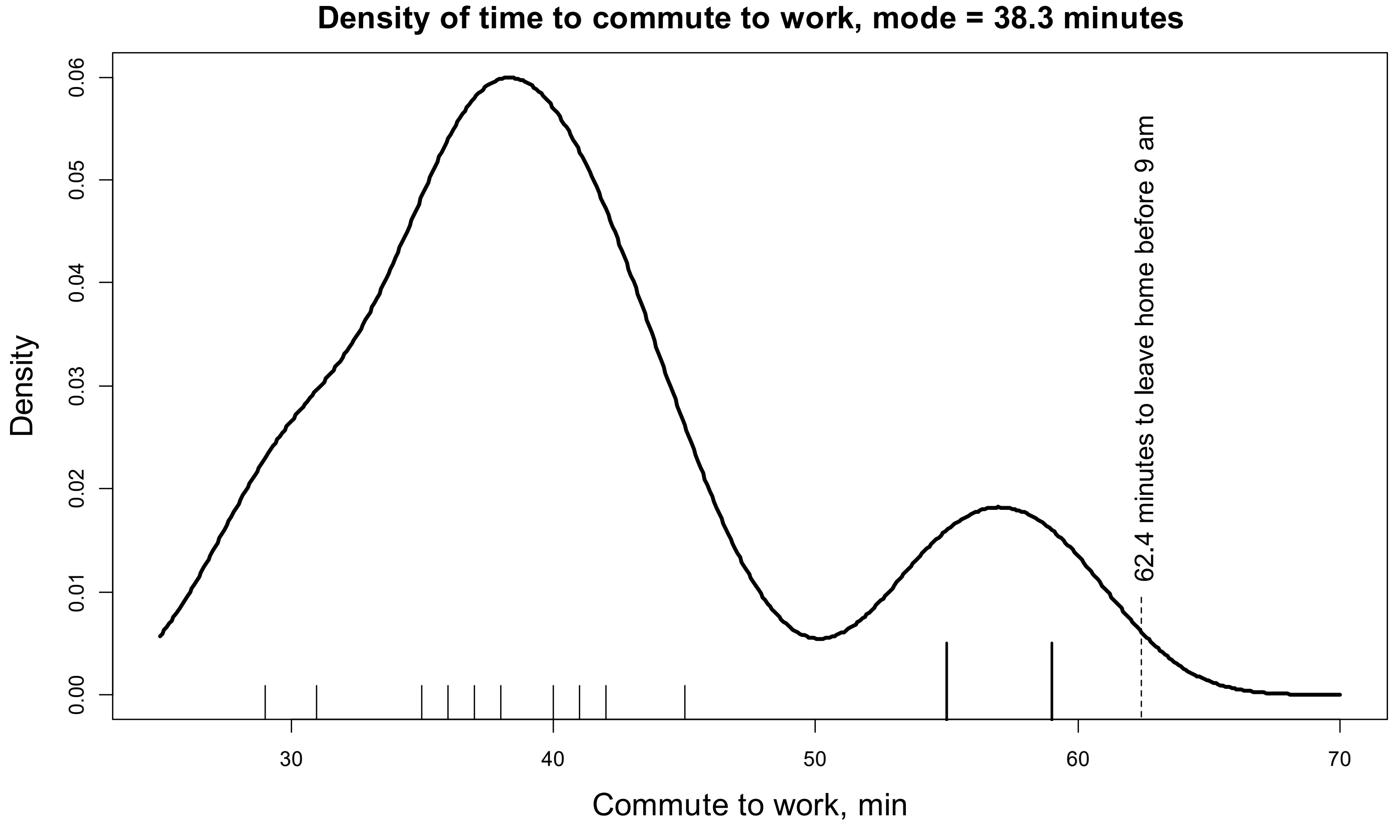
Example 7 ***Don't be late to the meeting.*** *Jack commutes to work via a busy highway. He recorded the time on 10 occasions (min): 38, 45, 31, 29, 35, 40, 37, 36, 42, 41. A couple of time there was a car accident and he spent 55 and 59 minutes in the traffic. One day Jack must give an important presentation at 9 AM and he wants to be on time with probability 0.99. When he needs to leave his home?*

Solution. First, we estimate the density of time commute using Gaussian kernel and second use Property 4 to derive the respective cdf. The two late times should not be excluded because we have to account for possible car accidents on the day of the meeting. Figure depicts the data and Gaussian kernel density to commute to work with default bandwidth (see the R function `jackM`). The time to leave home before 9 a.m. is the 99th percentile of the kernel cdf computed from the equation

$$\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{T_i - t}{h}\right) = 0.01,$$

where $T_i, i = 1, .., n = 12$ is the recorded travel time, including when car accidents happen, h is the `R` computed bandwidth (=2.86), and t is the 99th percentile.

How to find t? Computing the left-hand side on the grid of values and finding what t gives the closest value to 0.01. This gives $t = 62.4$ min. This means that Jack should leave home at 7:57 a.m. to be on time with probability 0.99.



Homework 1

Graphical presentation matters.

1. (10 points). Modify `toears` code to display the q-q plot for testing the normal and lognormal distribution along with the 95% beta q-q confidence band. Use `par(mfrow=c(1,2),mar=c(4.5,4.5,3,1), cex.lab=1.5 ,cex.main=1.5)`. Comment on the usefulness of the log transformation.

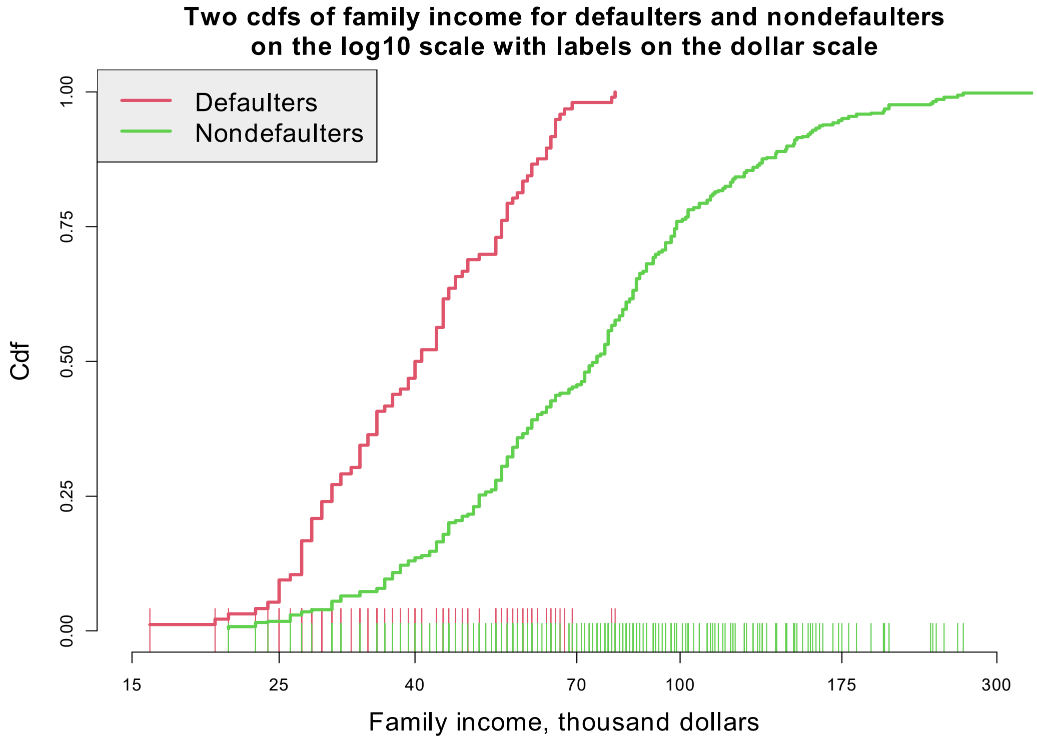
2. (10 points). File `height.csv` contains height (cm) of random people. Use `density` to reconstruct and plot the pdf (use **`rug`** command to show the data). Explain the result. Estimate the number of people taller than 250 cm in town with population 100,000. Display the result on the graph. [Hint: See "Don't be late to the meeting" Example and `jackM` code.]

3. (10 points). Compute three central tendency measures for original and log-transformed `toears` (don't forget to exponentiate the log-transformed centers). Explain the results: why some are different and some are close. Print out as the data frame with two columns and three rows. [Hint: Consult Section 2.11. Use *arithmetic* and *geometric* mean and their inequality, explain why the medians are the same.]

Solutions

See the R code `hw21_1` (listed at the end).

1. `hw21_1(job=1)`



Obviously, the original data do not follow normal distribution. The log transformation radically improves the normality although still not perfect.

2. The density plot hints to a mixture of two distributions: the distribution with the mode close to 152 cm is the distribution of female heights and the distribution with the mode close to 170 cm is the distribution of male heights. Since the kernel density is Gaussian the cdf is the average of normal cdfs, that is, similarly to the **Don't be late to the meeting** Example, the kernel cdf is a linear combination of normal cdfs:

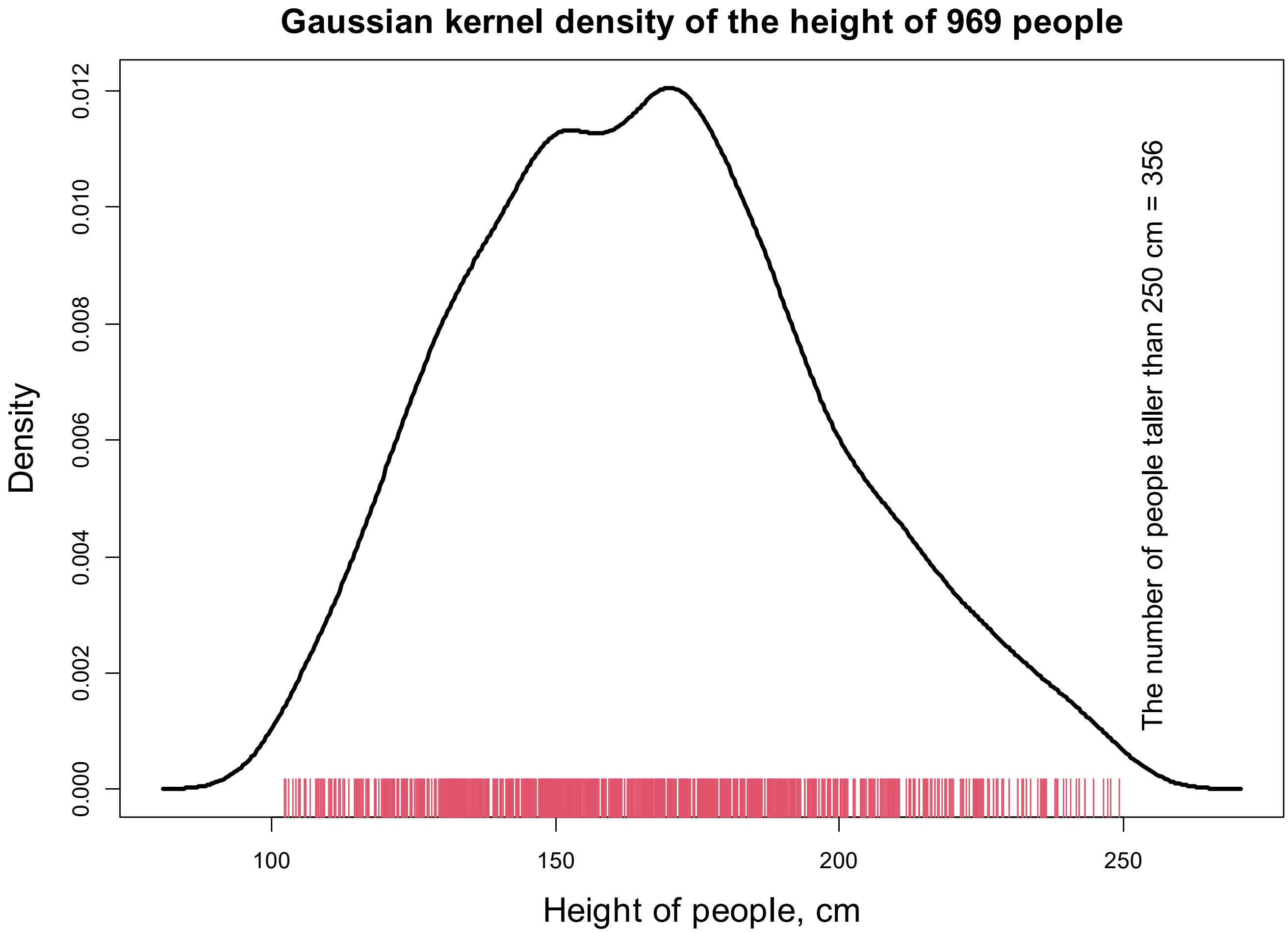
$$F(t) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{t-T_i}{h}\right),$$

the proportion of people who is shorter than t. The proportion of people who are taller than 250 is

$$1 - F(250) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{T_i-250}{h}\right) = 0.003564529$$

because $1-\Phi(x) = \Phi(-x)$. Thus, we estimate that the number of people taller than 250 cm is 356.

```
hw21_1(job=2)
```



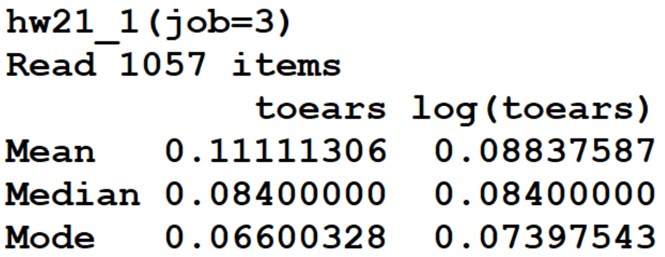
3. The mean of the original data is larger than exp mean of the log-transformed data, the geometric mean, because of the heavy right tail. See more discussion on the lognormal distribution in Section 2.11.

Column comparison. The median of the original and log-transformed is the same as is always the case. This is because the median is such that 50% of the data points are to the left and therefore any increasing monotonic transformation, such as log, preserves this property, shortly, median does not change. The mean derived from the log-transformed is the geometric mean:

$$e^{\frac{1}{n}\sum_{i=1}^{n}\ln X_i} = \left(e^{\sum_{i=1}^{n}\ln X_i}\right)^{1/n} = \left(\prod_{i=1}^{n} e^{\ln X_i}\right)^{1/n} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

because $e^{\ln a} = a$. It is known the the arithmetic mean > geometric mean (you can prove using Jensen's inequality). Also for a lognormal distribution geometric mean, `Mean log(toears)=0.08837587`, is close to the median `0.08440000` as follows from Section 2.11.

Row comparison. (1) For the first column with original data, when the distribution has a heavy right tail the mean is heavily influenced by large-value observations and therefore is the largest among the three centers (0.1111>0.0840 and 0.1111>0.066). Again, for a heavy right tail mode < median (0.084 > 0.066). (2) For the second column, the same order remains but with much less difference. It hints that, the log transformation does not completely eliminates the heavy right tail. Indeed, this can be seen from the q-q plot of log10 toears.



```
hw21_1(job=3)
Read 1057 items
           toears log(toears)
Mean   0.11111306  0.08837587
Median 0.08400000  0.08400000
Mode   0.06600328  0.07397543
```

```
hw21_1=function(job=1,lambda=0.95)
{
dump("hw21_1","c:\\QBS124\\hw21_1.r")
if(job==1)
{
par(mfrow=c(1,2),mar=c(4.5,4.5,3,1),cex.lab=1.5,cex.main=1.5)
x=scan("c:\\QBS124\\toears.txt");n=length(x)
ii=1:n;thq=qnorm((1:n)/n)
xst=(x-mean(x))/sd(x)
qn=qnorm((1:n)/n)
plot(qn,xst,,xlim=c(-3,3),ylim=c(-3,3),xlab="Theoretical quantile, Z",
        ylab="Sorted original toears",main="q-q plot for the original toears")
segments(-5,-5,5,5,col=2)
upB=qnorm(qbeta(.5+lambda/2,shape1=ii,shape2=n-ii+1))
lowB=qnorm(qbeta(.5-lambda/2,shape1=ii,shape2=n-ii+1))
lines(thq,upB,type="s",col=3)
lines(thq,lowB,type="s",col=3)
x10=log10(x)
x10=(x10-mean(x10))/sd(x10)
plot(qn,x10,xlim=c(-3,3),ylim=c(-3,3),xlab="Theoretical quantile, Z",
        ylab="Empirical quantile",main="q-q plot for normalized log10 toears")
segments(-5,-5,5,5,col=2)
upB=qnorm(qbeta(.5+lambda/2,shape1=ii,shape2=n-ii+1))
lowB=qnorm(qbeta(.5-lambda/2,shape1=ii,shape2=n-ii+1))
lines(thq,upB,type="s",col=3)
lines(thq,lowB,type="s",col=3)
}
if(job==2)
{
par(mfrow=c(1,1),mar=c(4.5,4.5,3,1),cex.lab=1.5,cex.main=1.5)
h=read.csv("c:\\QBS124\\height1000.csv")$height_cm
n=length(h)
dh=density(h)
plot(dh$x,dh$y,type="l",lwd=3,xlab="Height of people, cm",ylab="Density",
        main=paste("Gaussian kernel density of the height of",n,"people"))
rug(h,ticksize=0.05,col=2)
bw=dh$bw
pr250=mean(pnorm((h-250)/bw))
print(pr250)
text(255,0.001,paste("The number of people taller than 250 cm =",round(100000*pr250)),
        adj=0,srt=90,cex=1.25)
segments(250,-1,250,1,col=3)
}
if(job==3)
{
x=scan("c:\\QBS124\\toears.txt");n=length(x)
Lx=log(x)
da=data.frame(matrix(ncol=2,nrow=3))
names(da)=c("toears","log(toears)")
row.names(da)=c("Mean","Median","Mode")
da[1,1]=mean(x)
da[2,1]=median(x)
d=density(x)
da[3,1]=d$x[d$y==max(d$y)]
da[1,2]=exp(mean(Lx))
da[2,2]=exp(median(Lx))
d=density(Lx)
da[3,2]=exp(d$x[d$y==max(d$y)])
print(da)
}
}
```