Stochastic processes and random variables

Background

The probability model

Ott (1994)

A *stochastic process* is a process which includes or comprises random components. We describe the outcome of such processes with *random variables*, which can take on a range of values. A *probability model* is a set of rules describing the probabilities of all possible outcomes in the sample space. The values generated by such models are called *probability distributions*. We will discuss probability distributions for continuous random variables (environmental concentrations).

For a random variable X, we can describe probability ranges/distributions by the continuous distribution function (CDF) F_X and probability distribution function (PDF) f_X:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

$$f_X(x) = \frac{dP}{dx} = \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = \frac{d}{dx} F_X(x)$$

Any physical observation we make can be considered as “sampling” from this distribution. The actual value that we observe will depend on a large number of stochastic processes, but the likelihood of drawing a particular value will follow $f_x(x)$.

Appearance of randomness can arise from

- *variability*: natural variations

- *uncertainty*: "incomplete scientific or technical knowledge" (Morgan, Henrion, and Small 1992), or our lack of capability for accurate/precise observation. (our ignorance regarding functional dependences among variables may lead to the appearance of randomness)

Characterizing random variables with probability distributions, expected value, and variance

In this lesson, we will introduce nonparameteric and three parameteric distributions: uniform, normal, and lognormal distributions.

In addition, we will discuss two main properties of random variables.

- The expected value (also: average or arithmetic mean) of a random variable $\mathrm{E}(X)$ is a measure of the central tendency.
- The variance $\mathrm{Var}(X)$ is the second moment about the mean, $\mathrm{E}\{[X - \mathrm{E}(X)]^2\} = \mathrm{E}(X^2) - \mathrm{E}(X)^2$.

When we describe the probability model of a random variable X with a parameteric distribution, we can express $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ as a function of distribution parameters. The sample mean, $\bar{X}$, is also a random variable.

Parametric distributions

Uniform distribution

Random variable X can take on any value between a and b with equal probability.

For $x \in [a, b]$, the PDF and CDF are

$$f_X(x) = \frac{1}{b-a}$$
$$F_X(x) = \frac{x-a}{b-a}$$

Mean and variance:

$$\mathrm{E}(X) = \frac{a+b}{2}$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

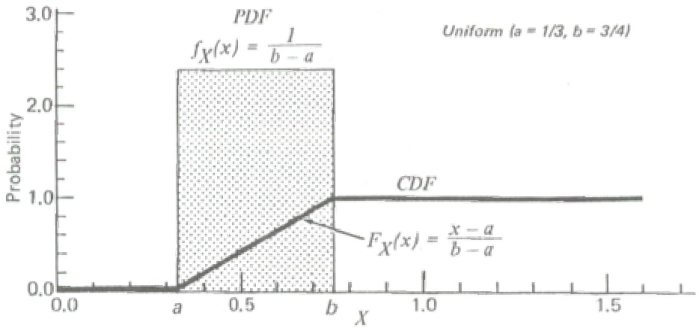


Figure 3.4 from Ott (1994)

Normal distribution

PDF and CDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$F_X(x) = \frac{1}{2}\left[\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) + 1\right]$$

The following ratio is also called the normal standard variable; often designated as z:

$$z = \frac{x-\mu}{\sigma}$$

The value of z is also called the z-score.

Mean and variance:

$$\mathrm{E}(X) = \mu$$
$$\mathrm{Var}(X) = \sigma^2$$

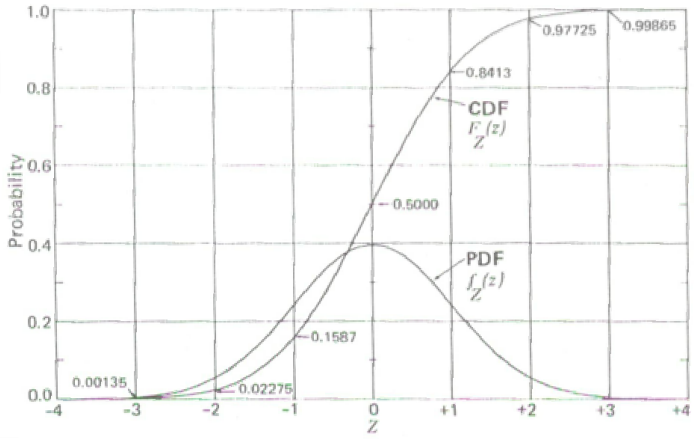


Figure 7.1 from Ott (1994)

Lognormal distribution

PDF and CDF:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$
$$F_X(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right]$$

Note that the expected value and variance of X are not the same μ and σ^2 as for the normal distribution.

$$\mathrm{E}(X) = e^{\mu+\sigma^2/2}$$
$$\mathrm{Var}(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

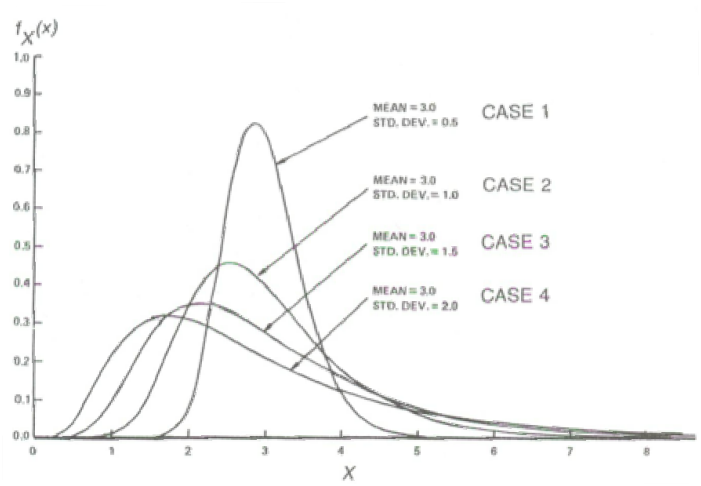


Figure 9.2 from Ott (1994)

Non-(globally)parameteric estimates of the PDF and CDF for a set of observations

For exploratory analysis, it is useful to visualize $f_X(x)$ and $F_X(x)$ that do not assume any fixed probability distributions a priori (empirical distributions).

PDF. A histogram normalized to probability rather than counts is an estimator of the density function. However, the histogram depends on the starting point of the grid of bins. Shown below are five histograms (of Old Faithful geyser duration data) with grid bins shifted (fixed width of 0.5) and the frequency polygon of their average:

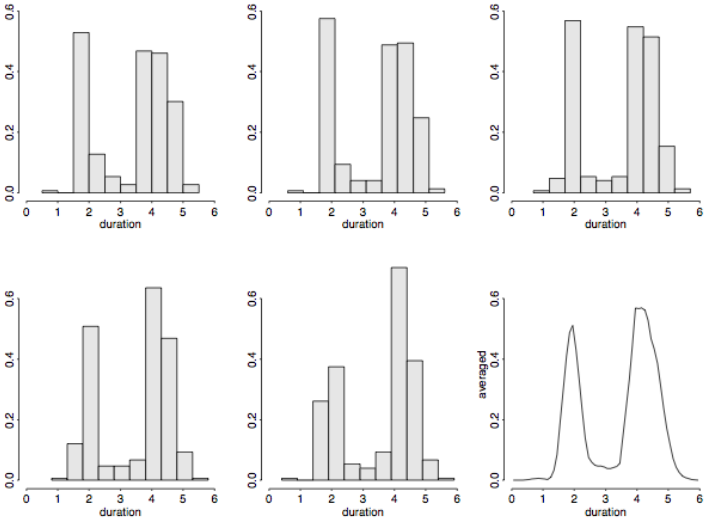


Figure 5.8 from Venables and Ripley (2003)

We can alternately view the *kernel density estimate* with bandwidth b,

$$\hat{f}_X(x) = \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - x_j}{b}\right)$$

The fixed kernel K is normally chosen to be a probability density function (PDF). For instance, for a normal (Gaussian) distribution is

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

Shown below are the same Old Faithful duration histograms superimposed by density estimates using a Gaussian kernel and several different bandwidths.

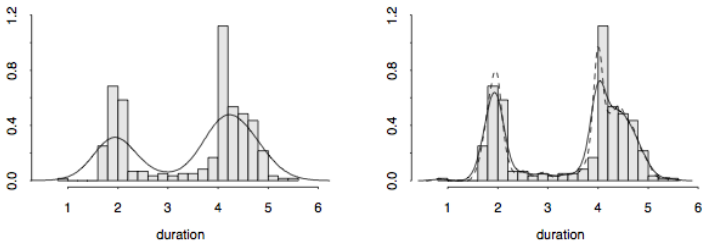


Figure 5.9 from Venables and Ripley (2003)

Note that the form of K is not related to the actual distribution of X. Alternative kernels include rectangular, triangular, cosine, etc. For instance, the cosine kernel would be $K(u) = (1 + \cos \pi u)/2$ over [-1,1].

Empiricial CDF. We can estimate the CDF by number of elements in sample set less than or equal to a value x divided by the number of elements in the sample set:

$$\hat{F}_X(x) = \frac{1}{n}\left|\{x_i \leq x \ : \ i = 1, 2, \ldots, n\}\right|$$

where $\left|\cdot\right|$ is the cardinality of (number of elements in) a set.

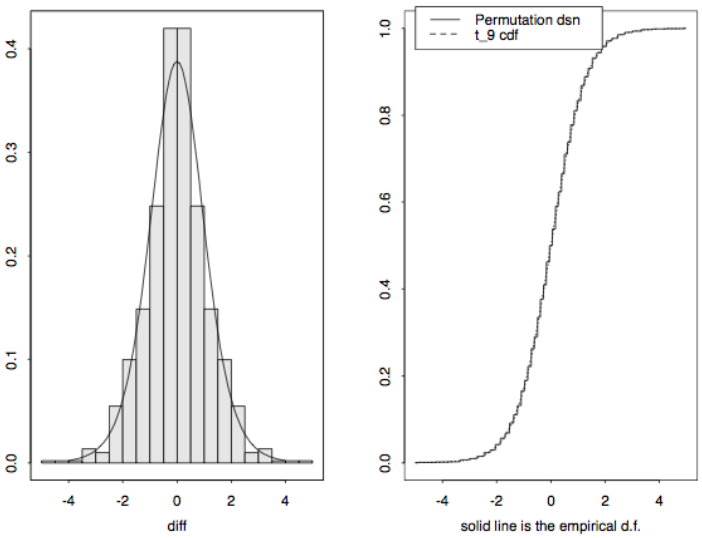


Figure 5.5 from Venables and Ripley (2003)

Types of statistics

We distinguish among three types of statistics:

- *Descriptive statistics*: Provide characterization of your random variables with meaningful metrics that simplify their representation. [covered in this lesson]
- *Inferential statistics*: Determine underlying mechanisms that generated the data, or draw conclusions regarding the value of metrics or estimates observed given the variability in the data. [covered in a future lesson]

- *Predictive statistics*: Build models to make forecasts (interpolation and extrapolation). Descriptive statistics of current or local data are often used to build such models.

Descriptive statistics

Probability distributions can be used to describe natural variability and uncertainties in our measurement or knowledge:

- Probability distribution can be used to describe *natural variability* of pollutant (e.g., NO_2) concentrations (due to mechanistic reasons). We can obtain more restrictive probability distributions by identifying the underlying factors and creating a number of probability distributions, each of which applies to a set of specific conditions.
- In many circumstances, we can describe probability distributions with a parameteric distribution (i.e., an analytical function with a fixed set of parameters). E.g., *a lognormal distribution*.
- We can imagine that there are *uncertainties* regarding our knowledge of the exact probability distribution of pollutant (e.g., NO_2). If we describe this distribution in parameteric form, then we can understand one aspect of the uncertainties in distribution through studying the uncertainties in the parameters of this distribution.
- Uncertainty regarding a parameter can also be characterized with a probability distribution of the estimator for that parameter.

As described above, characteristics most commonly used statistics to describe distributions of random variable are:

- expected value, written as $\mathrm{E}(X)$ or μ_X
- variance, written as $\mathrm{Var}(X)$ or σ_X^2

For a normal distribution, $\mu \pm \sigma$ encompasses 68.2%, $\mu \pm 2\sigma$ encompasses 95.5%, $\mu \pm 3\sigma$ encompasses 99.7%, and so on. We can also consider their geometric corollaries, which are appropriate for lognormally distributed variables: geometric mean ($\mu_{\log(X)}$) and geometric standard deviation ($\sigma_{\log(X)}$).

The mean and variance are examples of the possible *moments* of a distribution $f(x)$, which include skewness (3rd order), kurtosis (4th order), and so on.

$$M_n(X) = \int_{-\infty}^{\infty} (x-c)^n f(x) dx$$

Higher order moments ($n > 2$) are not commonly used in practice.

We can also compute *order statistics*, which do not assume any parameteric form $f_\theta(x)$ of the underlying distribution.

- median
- quartiles and interquartile-range (IQR)
- quantiles (e.g., 5th to 95th percentile)
- range

These metrics more generally describe the *central tendencies* and *dispersion* in the distribution. The difference between the mean and median may provide indication regarding the symmetry of the distribution.

R Demonstration

```
library(dplyr)
library(reshape2)
library(chron)
library(ggplot2)
```

```
source("GRB001.R")
```

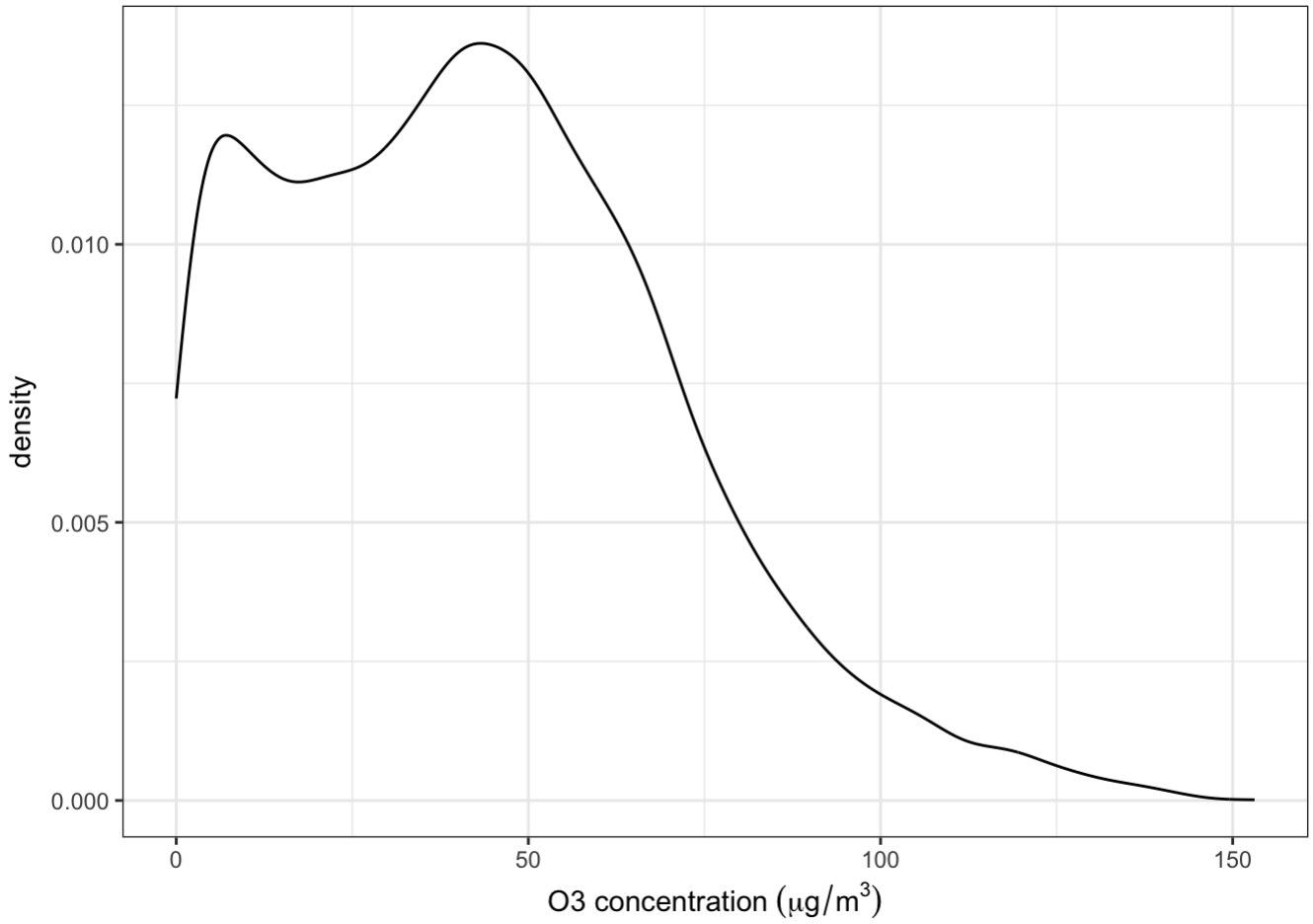
```
Sys.setlocale("LC_TIME","C")
options(stringsAsFactors=FALSE)
options(chron.year.abb=FALSE)
theme_set(theme_bw())
```

```
df <- readRDS("data/2013/lau-zue.rds")
lau <- df %>% filter(site=="LAU")
```

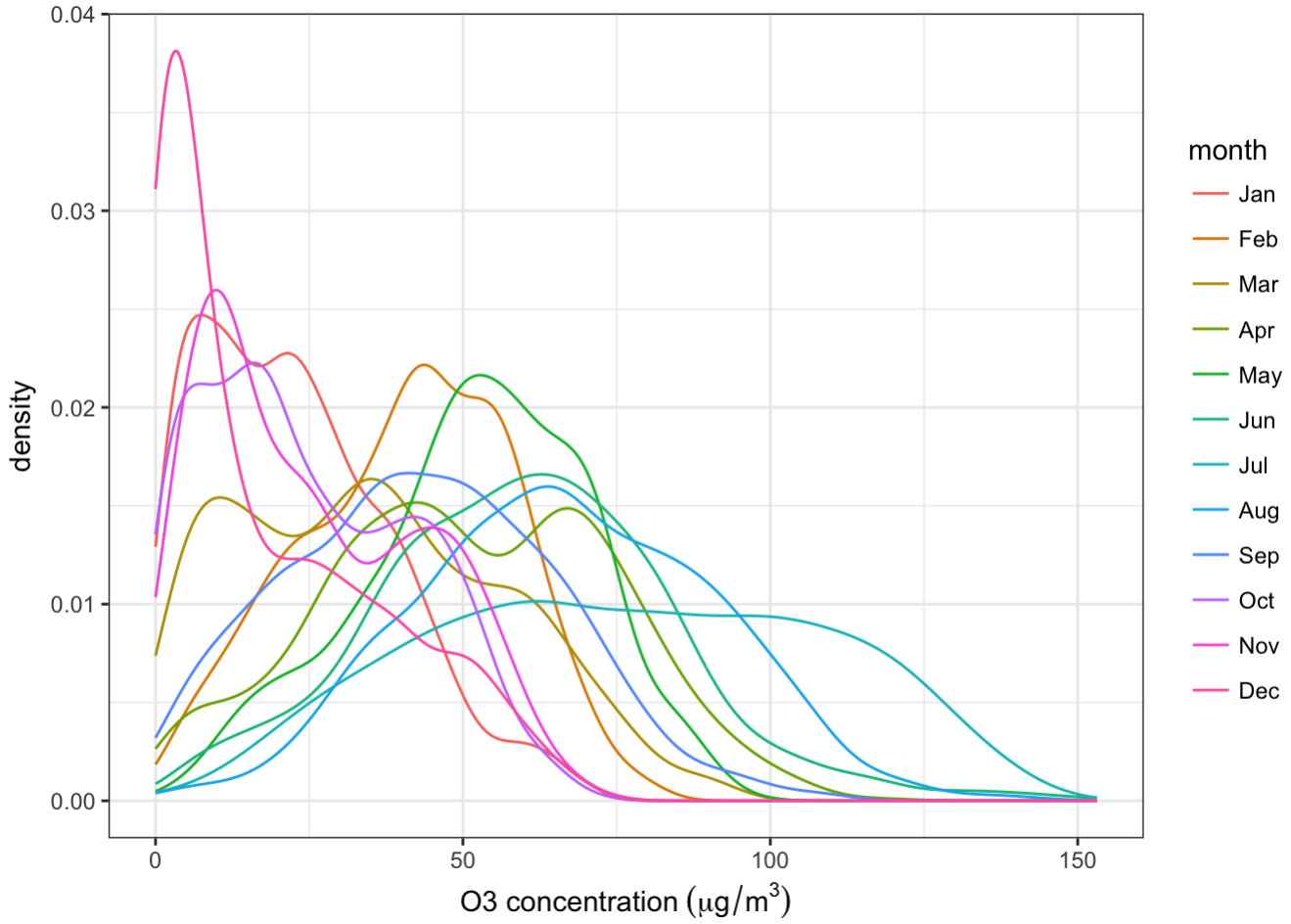
Ozone in Lausanne (probability/kernel density)

The annual probability density appears to be multimodal, but we can see that this is due to the strong variation in seasonal probability distributions of ozone concentrations.

```
ggp <- ggplot(lau)+
  geom_line(aes(x=O3, ymax=..density..),
            stat="density", position="identity")+
              xlab(expression("O3 concentration"~(mu*g/m^3)))
print(ggp)
```



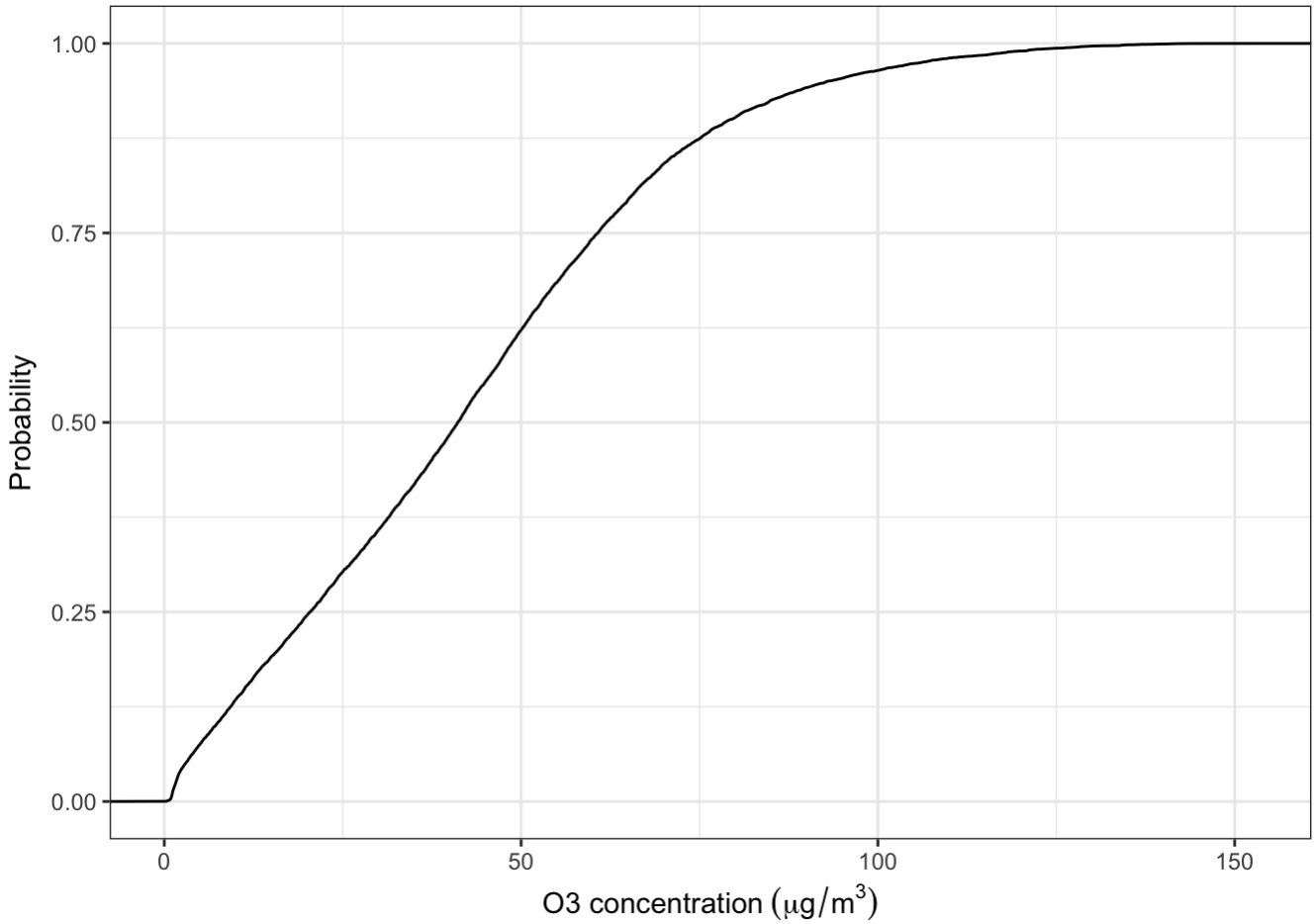

```
ggp <- ggplot(lau)+
  geom_line(aes(x=O3, ymax=..density.., group=month, color=month),
            stat="density",
            position="identity")+
              xlab(expression("O3 concentration"~(mu*g/m^3)))
print(ggp)
```



Ozone in Lausanne (ECDF)

We can also view the empirical CDFs (ECDF) of hourly concentrations for 2013. ECDFs can be computed with R in (see `?ecdf`), but can be applied during the call to `ggplot`.

```
ggp <- ggplot(lau)+
  geom_line(aes(x=O3), stat="ecdf")+
    xlab(expression("O3 concentration"~(mu*g/m^3)))+
      ylab("Probability")
print(ggp)
```



We can also segregate the ECDFs by month; for ozone there is a particularly strong variation.

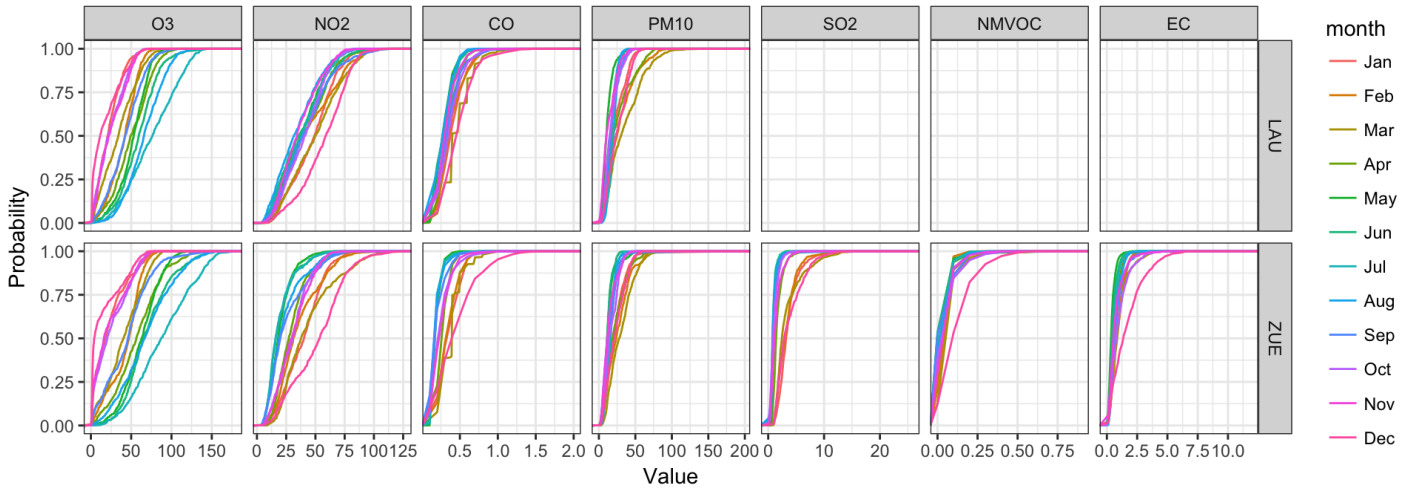
```
ggp <- ggplot(lau)+
  geom_line(aes(x=O3, group=month, color=month),
            stat="ecdf", position="identity")+
  xlab(expression("O3 concentration"~(mu*g/m^3)))+
  ylab("Probability")
```

All pollutants in Lausanne and Zurich

First, we must convert our data frame to the elongated format.

```
id.vars <- c("datetime", "season", "month", "site")
pollutants <- c("O3", "NO2", "CO", "PM10", "SO2", "NMVOC", "EC")
lf <- melt(df[, c(id.vars, pollutants)], id.vars=id.vars)
```

```
ggp <- ggplot(lf)+
  geom_line(aes(x=value, group=month, color=month),
            stat="ecdf", position="identity")+
  facet_grid(site~variable,scale="free_x")+
  xlab("Value")+
  ylab("Probability")
print(ggp)
```



Descriptive statistics

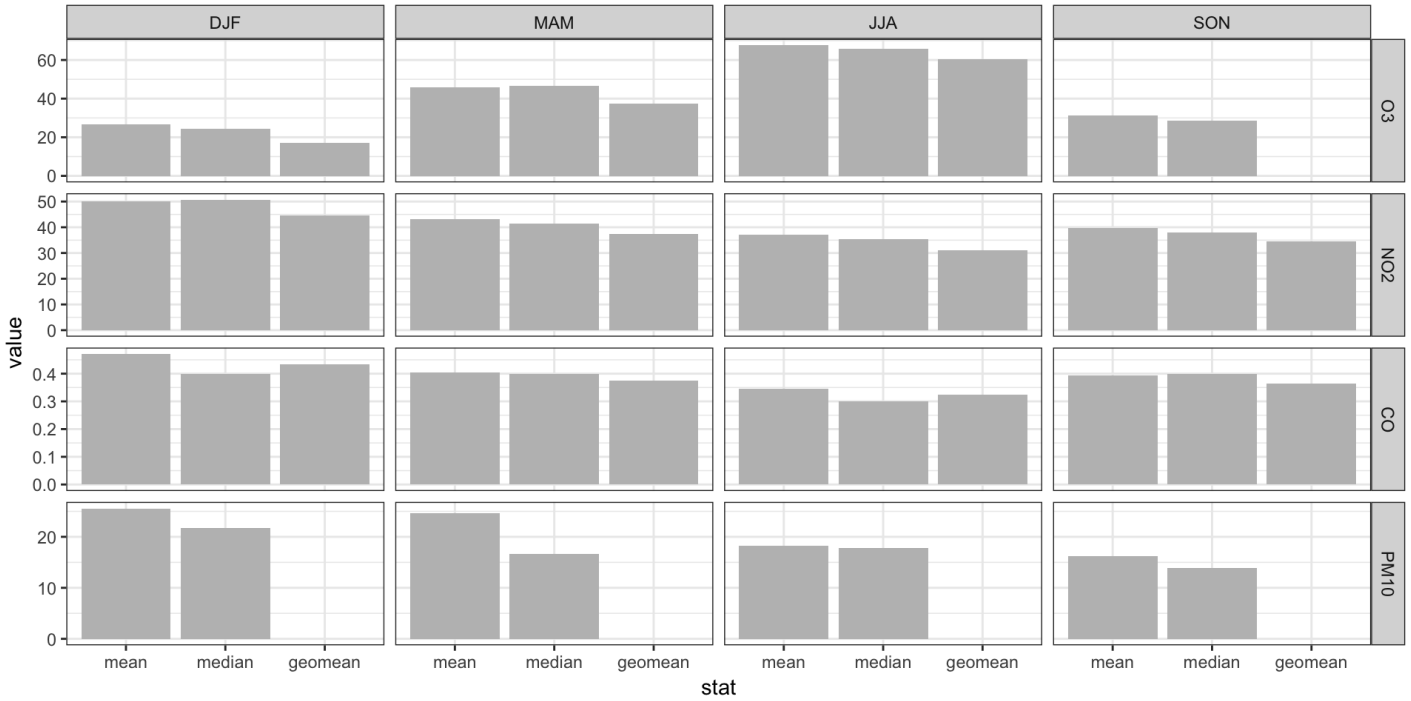
Define some functions:

```
Geomean <- function(x, na.rm = FALSE, trim = 0, ...) {
  ## Geometric mean
  exp(mean(log(x, ...), na.rm = na.rm, trim = trim, ...))
}

Geosd <- function(x, na.rm = FALSE, ...) {
  ## Geometric standard deviation
  exp(sd(log(x, ...), na.rm = na.rm, ...))
}

ComputeCentral <- function(x) {
  ## x is a vector of values
  stats <- c("mean"    = mean(x,na.rm=TRUE),
             "median"  = median(x,na.rm=TRUE),
             "geomean" = Geomean(x,na.rm=TRUE))
  data.frame(stat=factor(names(stats), names(stats)), value=stats)
}

ComputeDispersion <- function(x) {
  ## x is a vector of values
  stats <- c("sd"    = sd(x,na.rm=TRUE),
             "IQR"   = IQR(x,na.rm=TRUE),
             "geosd" = Geosd(x,na.rm=TRUE))
  data.frame(stat=factor(names(stats), names(stats)), value=stats)
}
```

Compare central values:

```
stats.central <- lf %>% filter(site=="LAU" & !is.na(value)) %>%
  group_by(season, variable) %>%
  do(ComputeCentral(.[["value"]]))
```

```
ggp <- ggplot(stats.central)+
  geom_bar(aes(stat, value),
           stat="identity",
           fill="gray")+
  facet_grid(variable~season,scale="free_y")
print(ggp)
```

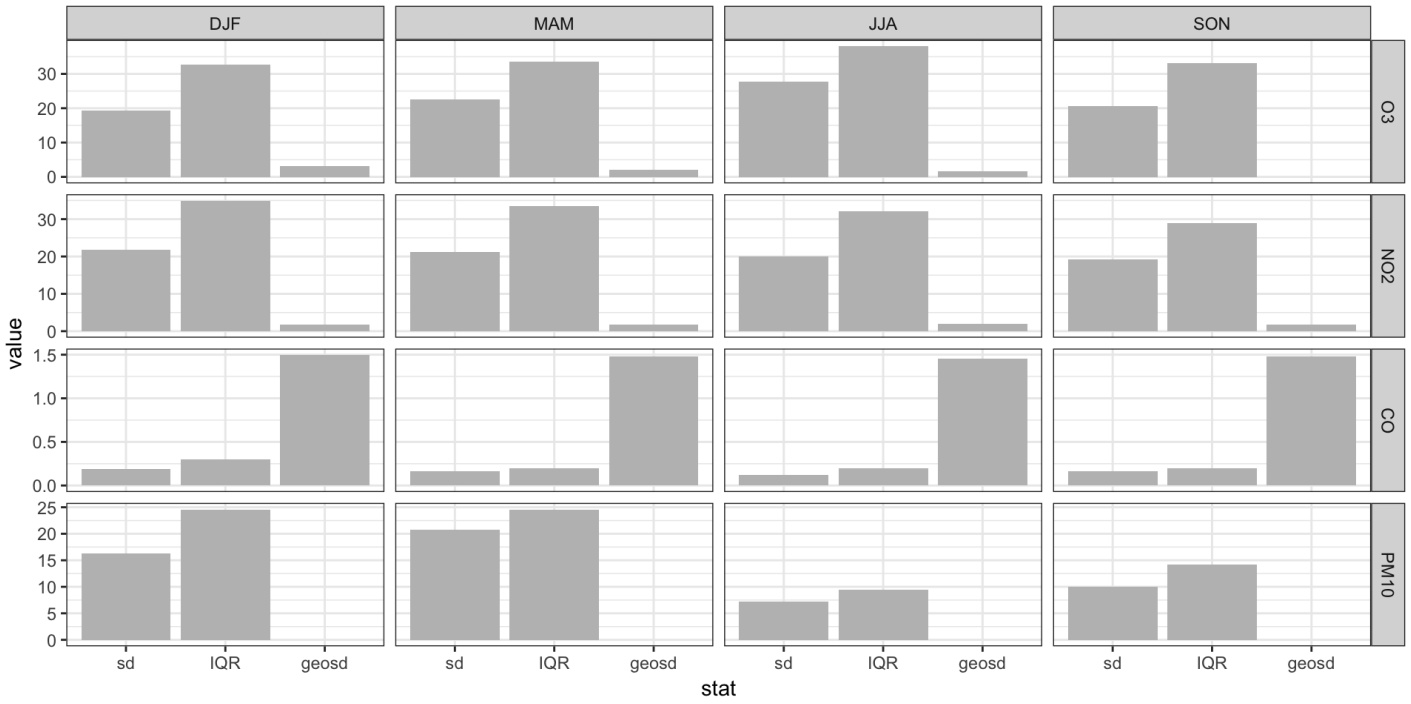


Why are geomean values missing?

Compare dispersion metrics:

```
stats.disp <- lf %>% filter(site=="LAU" & !is.na(value)) %>%
  group_by(season, variable) %>%
  do(ComputeDispersion(.[["value"]]))
```

```
ggp <- ggplot(stats.disp)+
  geom_bar(aes(stat, value),
           stat="identity",
           fill="gray")+
  facet_grid(variable~season,scale="free_y")
print(ggp)
```



Why are geosd values missing?

References

Morgan, M.G., M. Henrion, and M. Small. 1992. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press.

Ott, Wayne R. 1994. *Environmental Statistics and Data Analysis*. Taylor & Francis.

Venables, W. N., and B. D. Ripley. 2003. *Modern Applied Statistics with S*. Springer.