

Introduction to Geostatistics

5. Probability II: random variables and probability distributions

Edzer Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (ifgi)
University of Münster

May 11, 2010



Random variables

- ▶ A random variable is a numerical variable whose outcome is subject to chance.
- ▶ We say (loosely) that the probability of taking on a certain value is denoted by its *probability density* $p(x)$ (or $f(x)$).
- ▶ Examples of discrete variables: throwing a dice, tossing a coin
- ▶ Examples of continuous variables: *exact* body length of a randomly sampled person
- ▶ If outcomes were completely predictable, there would be no element of chance
- ▶ How can we describe probability distributions?
- ▶ Discrete random variables, continuous random variables



Probability density and distribution function

Probability *density* $f(x)$ gives, for discrete variables, the amount of probability of being x , and is non-negative: $f(x) = Pr(X = x)$

Probability *distribution* ranges from 0 to 1, and gives the cumulative probability up to x .

For discrete variables

$$F(x_i) = \sum_{x \leq x_i} f(x)$$

for continuous variables

$$F(x_i) = \int_{-\infty}^{x_i} f(x) dx$$



Expectation, Variance

Expectation is the mean value for a random variable. Discrete RV:

$$E(X) = \mu = \sum_{i=1}^n x_i f(x_i)$$

Continuous RV:

$$E(X) = \mu = \int_{-\infty}^{+\infty} xf(x) dx$$

note that $E(X)$ is a numeric value, i.e. is non-random, and that the argument of $E(\cdot)$ is random.

How does the expectation of X relate to the sample mean,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i?$$

For *random* sampling a sample of size n from an infinite population, each $f(x_i)$ is estimated by $\frac{1}{n}$, and $\hat{\mu} = \bar{x}$.



Variance, covariance

variance of a random variable is defined in terms of expectation

$$\text{Var}(X) = E(X - E(X))^2$$

covariance is a measure of co-variation of two random variables

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

with the following properties:

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$$

if X and Y are stochastically independent, then $\text{Cov}(X, Y) = 0$
(the reverse does not hold)



Covariance, correlation

If

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

and

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$$

then the *correlation coefficient* between X and Y ,

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

must have the property that

$$-1 \leq r(X, Y) \leq 1$$

it measures strength of *linear* relationship, and is 0 in absence of a linear relation, 1 (-1) if the relationship is perfect, ascending (descending).



Moments

The k -th moment of X is defined as

$$\mu'_k = E(X^k)$$

The k -th central moment of X is defined as

$$\mu_k = E((X - E(X))^k)$$

One can define a probability density function by all its moments. The third central moment is of interest, as it tells whether a distribution is symmetric ($\mu_3 = 0$), or *skew*. Is it right-skew, then $\mu_3 > 0$, is it left-skew then $\mu_3 < 0$.



Bernoulli distribution

$$X = \begin{cases} 1 & \text{red ball} \\ 0 & \text{blue ball} \end{cases}$$

$$f(k) = \begin{cases} q = 1 - p & \text{for } k = 0 \\ p & \text{for } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

p is the probability of success (1), q the probability of failure.



Binomial distribution

From n independent observations of a Bernoulli process having success probability p , we obtain exactly k hits with probability

$$f(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{(n-k)} & \text{for } k = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

example: random drawing n balls from a bowl, with replacement; what is the probability of drawing exactly k red balls.

```
> pbinom(4, 9, 0.5)
```

```
[1] 0.5
```

```
> dbinom(4, 9, 0.5)
```

```
[1] 0.2460938
```

```
> rbinom(20, 9, 0.5)
```

```
[1] 6 7 4 5 5 3 7 2 2 6 4 5 5 6 2 4 3 6 4 7
```



Poisson distribution

Special case of the Binomial distribution, where $n \rightarrow \infty$, and the rate $np = \lambda$ is known. The Poisson distribution describes the expected frequencies of "hits": $Pr(X = x|\lambda) = f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

Examples:

- ▶ length of a row in a shop (queueing problem)
- ▶ discrete events in temporal processes, spatial processes
- ▶ usually the Poisson is the base-line case, against which more structured processes are investigated



Uniform distribution

the continuous uniform distribution has a uniform density between its minimum value a and maximum value b :

$$f(x) = \begin{cases} 1/(b-a) & \text{for } a < x < b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

```
> runif(n = 10, min = 25, max = 50)
```

```
[1] 35.71244 49.27612 41.59993 30.49697 32.38869 47.24659 47.25337  
[8] 26.49742 39.28409 48.85743
```

```
> punif(9, min = 0, max = 10)
```

```
[1] 0.9
```



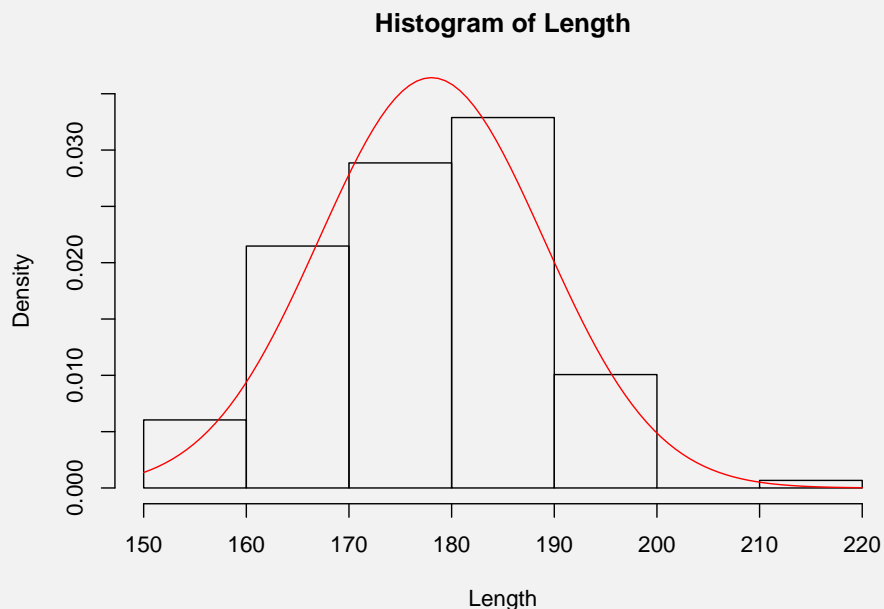
Normal (Gaussian) distribution

```
> m = mean(Length)
```

```
> s = sqrt(var(Length))
```

```
> hist(Length, probability = TRUE, ylim = c(0, 0.035))
```

```
> curve(dnorm(x, m, s), add = TRUE, col = "red")
```



Gaussian density function

No need to remember this:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

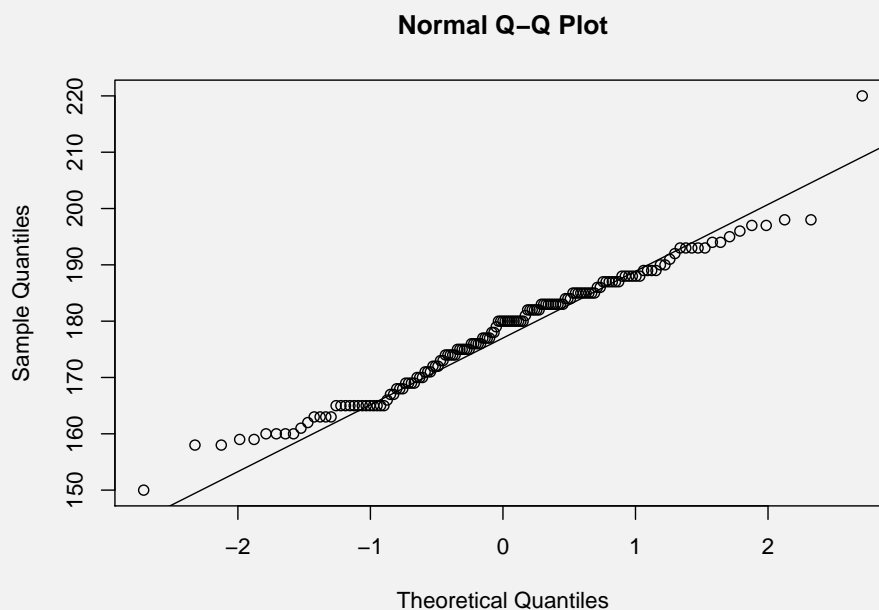
But remember:

- ▶ *only* depends on mean μ and standard deviation σ
- ▶ mean μ is also median: symmetric
- ▶ ranges from $-\infty$ to ∞
- ▶ approx. 68% lies between $\mu - \sigma$ and $\mu + \sigma$
- ▶ approx. 95% lies between $\mu - 2\sigma$ and $\mu + 2\sigma$



Normal probability plot

- > `qqnorm(Length)`
- > `qqline(Length)`



Normal probability plot (2)

```
> qqnorm(Length)
> qqline(Length)
> qqline2 <- function(x) {
+   m = mean(x)
+   s = sd(x)
+   lines(cbind(c(-3, 3), c(m - 3 * s, m + 3 * s)), col = "red")
+ }
> qqline2(Length)
```

