Statistical Data Analysis, Lecture 2

dr. Dennis Dobler

Vrije Universiteit Amsterdam

February 10, 2021

Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 3: Exploring distributions

Contents of Chapter 3:

1. Quantile function
2. Location-scale family
3. QQ-plots and symplots
4. Goodness-of-fit tests
   - Shapiro-Wilk test
   - Kolmogorov-Smirnov test
   - Chi-square test

distributions

Distribution functions

Random variable $X : \Omega \to \mathbb{R}$

Population distribution function $F(x) = P(X \leq x), \quad x \in \mathbb{R}$

Empirical distribution function given sample $x_1, \ldots, x_n$:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{x_j \leq x\}}, \quad x \in \mathbb{R}$$

In R, sample stored in vector `x`:

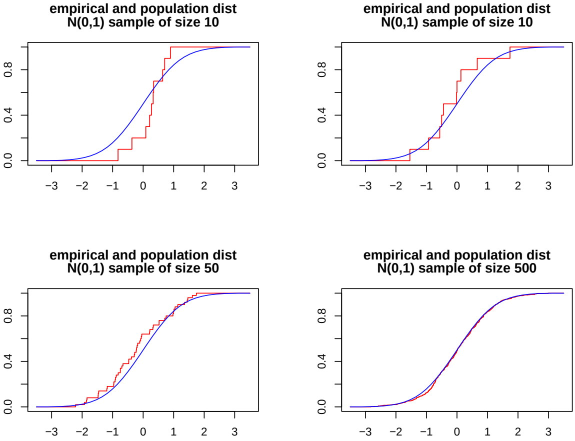
```
x <- rnorm(10, 0, 1)
plot(ecdf(x), col="red")
```

compare to e.g. $N(0,1)$ distribution function:

```
x.lattice <- seq(-3,3,0.001)
lines(x.lattice, pnorm(x.lattice,mean=0,sd=1), col="blue")
```

Example distribution functions

samples $x_1, \ldots, x_n$ from $N(0,1)$:



Goal (1)

Distribution: part of (parametric) statistical model.

Empirical distribution $\rightsquigarrow$ set up statistical model.

Goal: find underlying distribution

Goal (2)

Questions:

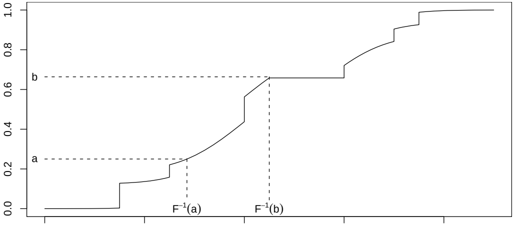
- One sample data from
  - specific distribution? (QQ-plot, goodness-of-fit tests)
  - symmetric distribution? (symplot)
- Two sample data from
  - same distribution? (QQ-plot)

quantile function and location-scale family

Quantile function

Quantile function F^{-1}: "inverse" of F.

Definition $F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}, \quad \alpha \in (0, 1).$



R: qnorm, qexp, qpois, etc.

Definition location-scale family

If $X \sim F$, denote $Y = a + bX \sim F_{a,b}, \quad a \in \mathbb{R},\ b > 0.$

$$\Rightarrow \quad F_{a,b}(x) = F\left(\frac{x-a}{b}\right)$$

Location-scale family (LSF) w.r.t. F: $\quad \{F_{a,b} : a \in \mathbb{R}, b > 0\}$

$\mathrm{E}(X) = 0\ \ \&\ \ \mathrm{var}(X) = 1 \Longrightarrow \mathrm{E}(Y) = a\ \ \&\ \ \mathrm{var}(Y) = b^2.$

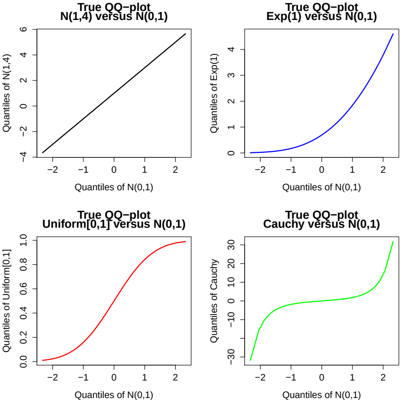
Quantiles of F and $F_{a,b}$ (1)

Claim $F_{a,b}^{-1}(\alpha) = a + bF^{-1}(\alpha)$.

Proof (for invertible F):

Quantiles of F and $F_{a,b}$ (2)

Hence, $\{(F^{-1}(\alpha), F_{a,b}^{-1}(\alpha)) : \alpha \in (0,1)\}$ forms line $y = a + bx$.



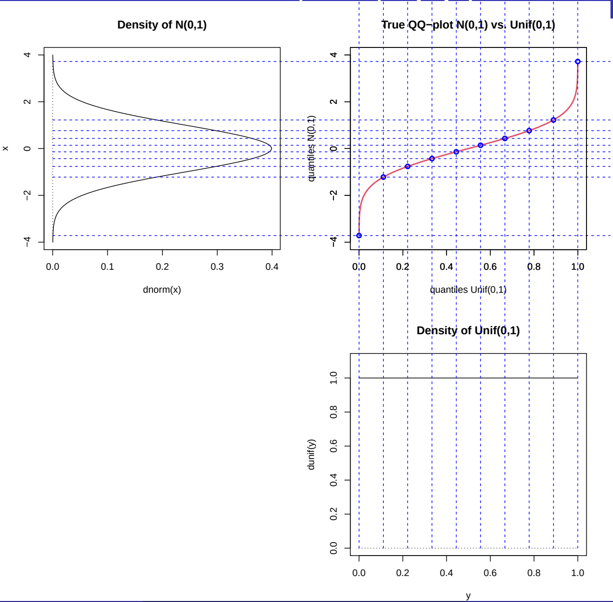
$N(0,1)$ & $N(1,4)$: same LSF.

$N(0,1)$ & $Exp(1)$: not same LSF.

$N(0,1)$ & $Unif(0,1)$: not same LSF.

$N(0,1)$ & $Cauchy(1)$: not same LSF.

Relative heaviness of tails in (true) QQ-plots



QQ-plots and symplots

Definition QQ-plot

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F \text{ continuous:} \quad \mathrm{E}F(X_{(i)}) = \tfrac{i}{n+1} \quad \Rightarrow \quad X_{(i)} \approx F^{-1}(\tfrac{i}{n+1}).$$

Definition QQ-plot

$X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ continuous: $\mathrm{E}F(X_{(i)}) = \frac{i}{n+1} \quad \Rightarrow \quad X_{(i)} \approx F^{-1}(\frac{i}{n+1})$.

If $Y_i = a + bX_i, \ i = 1, \dots, n:$ $\mathrm{E}F_{a,b}(Y_{(i)}) = \frac{i}{n+1}$.

Definition QQ-plot

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ continuous: $\mathrm{E}F(X_{(i)}) = \frac{i}{n+1} \quad \Rightarrow \quad X_{(i)} \approx F^{-1}(\frac{i}{n+1})$.

If $Y_i = a + bX_i,\ i = 1, \ldots, n:\ \mathrm{E}F_{a,b}(Y_{(i)}) = \frac{i}{n+1}$.

$\Rightarrow \quad Y_{(i)} \approx F_{a,b}^{-1}(\frac{i}{n+1}) = a + bF^{-1}(\frac{i}{n+1})$.

$\Rightarrow \ \left\{\left(F^{-1}\left(\frac{i}{n+1}\right), y_{(i)}\right) : i = 1, \ldots, n\right\}$ approximates line $y = a + bx$.

Definition QQ-plot

$X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ continuous: $\mathrm{E}F(X_{(i)}) = \frac{i}{n+1} \quad \Rightarrow \quad X_{(i)} \approx F^{-1}(\frac{i}{n+1})$.

If $Y_i = a + bX_i, \; i = 1, \dots, n$: $\mathrm{E}F_{a,b}(Y_{(i)}) = \frac{i}{n+1}$.

$\Rightarrow \quad Y_{(i)} \approx F_{a,b}^{-1}(\frac{i}{n+1}) = a + bF^{-1}(\frac{i}{n+1})$.

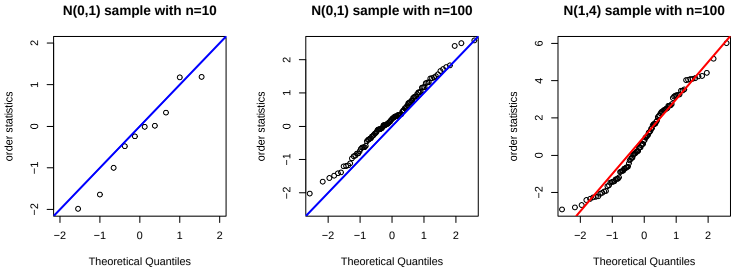
$\Rightarrow \; \left\{ \left(F^{-1}\left(\frac{i}{n+1}\right), y_{(i)} \right) : i = 1, \dots, n \right\}$ approximates line $y = a + bx$.

true F unknown... $\rightsquigarrow$ QQ-plots!

R: qqnorm, qqexp, qqunif, etc.

Example QQ-plot

of $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ vs $N(0,1)$; varying n, μ, σ^2.



$y = x$ $\quad$ $y = x$ $\quad$ $y = 1 + 2x$

In R:

```
par(pty="s")
qqnorm(rnorm(10))
```

Using QQ-plots — Example 1

- plot histogram
- plot different QQ-plots & choose most linear

Using QQ-plots — Example 1

- plot histogram
- plot different QQ-plots & choose most linear
- determine location (a) & scale (b) by fitting
  - straight line $y = a + bx$ visually
  - sample mean & variance
    to theoretical values (preferred!):
    $Y = a + bX \Rightarrow E(Y) = a + bE(X)$ & $var(Y) = b^2 var(X)$
    $\Rightarrow b = \sigma_Y / \sigma_X$ estimated by $\hat{b} = \hat{\sigma}_Y / \sigma_X$
    & $a = E(Y) - bE(X)$ estimated by $\hat{a} = \bar{Y} - \hat{b}E(X)$
    In R: use `abline(a = ...  , b = ...)` (different from qqline!)

Using QQ-plots — Example 1

- plot histogram
- plot different QQ-plots & choose most linear
- determine location (a) & scale (b) by fitting
  - straight line $y = a + bx$ visually
  - sample mean & variance

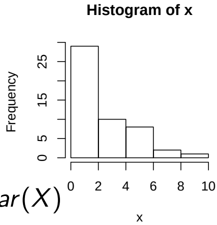
    to theoretical values (preferred!):

    $Y = a + bX \Rightarrow E(Y) = a + bE(X)$ & $var(Y) = b^2 var(X)$

    $\Rightarrow b = \sigma_Y / \sigma_X$ estimated by $\hat{b} = \hat{\sigma}_Y / \sigma_X$

    & $a = E(Y) - bE(X)$ estimated by $\hat{a} = \bar{Y} - \hat{b}E(X)$

    In R: use `abline(a = ... , b = ...)` (different from qqline!)



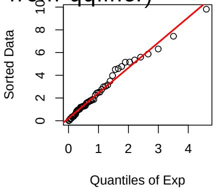
- Example $\bar{Y} = 1.98$, $\hat{\sigma}_Y^2 = 4.2$, $X \sim Exp(1)$

  $\Rightarrow$ $\hat{b} = 2.049$, $\hat{a} = 1.98 - \sqrt{4.2}/1 \approx -0.069$

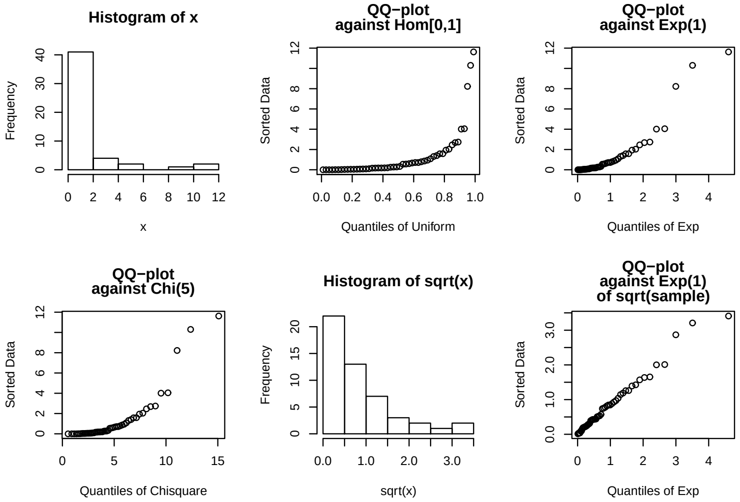
  $\Rightarrow$ model distribution $Y \sim -0.069 + 2.049 \cdot Exp(1)$

  $= -0.069 + Exp(1/2.049)$

  - Or round: $Y \sim Exp(0.5)$ (rather don't!)



Using QQ-plots — Example 2



Example: possibly transform.

Definition symmetry plot

Investigate symmetry/skewness of a distribution.

F symmetric around $\theta \;\Rightarrow\; F^{-1}(1-\alpha) - \theta = \theta - F^{-1}(\alpha), \quad \alpha \in (0,1).$

$\Rightarrow\; \{(\theta - F^{-1}(\alpha),\ F^{-1}(1-\alpha) - \theta) : \alpha \in (0,1)\}$ straight line $y = x$.

Definition symmetry plot

Investigate symmetry/skewness of a distribution.

F symmetric around $\theta \;\Rightarrow\; F^{-1}(1-\alpha) - \theta = \theta - F^{-1}(\alpha), \quad \alpha \in (0,1)$.

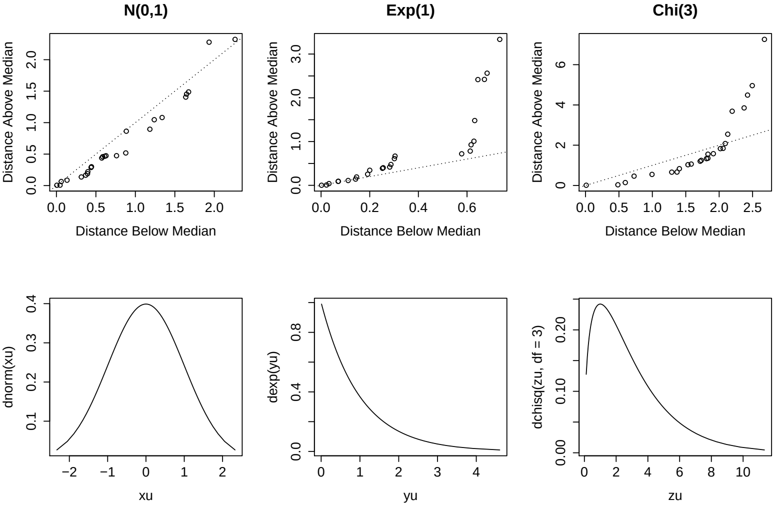
$\Rightarrow\; \{(\theta - F^{-1}(\alpha),\ F^{-1}(1-\alpha) - \theta) : \alpha \in (0,1)\}$ straight line $y = x$.

Symplot of $x_1, \ldots, x_n$: plot of

$$\left\{ (\mathrm{med}(x) - x_{(i)}, x_{(n-i+1)} - \mathrm{med}(x)) : i = 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor \right\}.$$

R: symplot

Example symplot



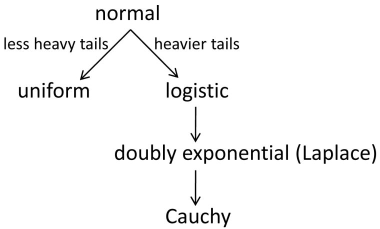
Sample size matters!

Other ways to investigate symmetry

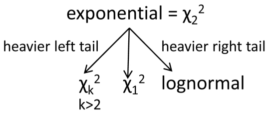
- histogram
- boxplot
- skewness parameter (be cautious, see syllabus)
- difference sample mean & sample median

Systematic search for underlying distribution

- Investigate symmetry plot & histogram
- Try several QQ-plots
  - if symmetric:



  - if not symmetric:



- Not satisfactory? Try transformations!

Two sample QQ-plot

$x_1, \ldots, x_m$ & $y_1, \ldots, y_n$: from same LSF?

If $m = n$: empirical QQ-plot plots $\{(x_{(i)}, y_{(i)}) : i = 1, 2, \ldots, n\}$.

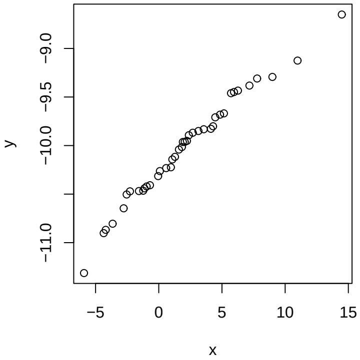
If $m < n$, it plots $\{(x_{(i)}, y^*_{(i)}) : i = 1, 2, \ldots, m\}$, where

$$y^*_{(i)} = \frac{1}{2}\left(y_{([i\frac{n+1}{m+1}])} + y_{([i\frac{n+1}{m+1} + \frac{m}{m+1}])}\right).$$

Idea: match $x_{(i)}$ with $y_{(j)}$ for which $\frac{i}{m+1} \approx \frac{j}{n+1}$.

R: qqplot

Two sample QQ-plot, example



Roughly straight line: possibly same LSF.

(GoF) tests

goodness-of-fit tests

Recap hypothesis testing:

- H_0, H_1, α,
- test statistic,
- its H_0-distribution,
- test score,
- p-value OR critical region,
- conclusion

Goodness-of-fit test

Idea: sample $x_1, \dots, x_n$ from unknown F. Test

$$H_0 : F \in \mathcal{F}_0$$
$$H_1 : F \notin \mathcal{F}_0$$

where $\mathcal{F}_0 = \{F_0\}$ (simple H_0)
or $\mathcal{F}_0 =$ collection of distributions (composite H_0), e.g. LSF.

Goodness-of-fit test

Idea: sample $x_1, \ldots, x_n$ from unknown F. Test

$$H_0 : F \in \mathcal{F}_0$$
$$H_1 : F \notin \mathcal{F}_0$$

where $\mathcal{F}_0 = \{F_0\}$ (simple H_0)
or $\mathcal{F}_0 =$ collection of distributions (composite H_0), e.g. LSF.

Aim: omnibus test with reasonable power.

Goodness-of-fit test

Idea: sample $x_1, \dots, x_n$ from unknown F. Test

$$H_0 : F \in \mathcal{F}_0$$
$$H_1 : F \notin \mathcal{F}_0$$

where $\mathcal{F}_0 = \{F_0\}$ (simple H_0)
or $\mathcal{F}_0 =$ collection of distributions (composite H_0), e.g. LSF.

Aim: omnibus test with reasonable power.

Interpretation: is H_0 not too implausible?

Different tests we consider

- Shapiro-Wilk: $H_0 : F \in \{N(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma^2 > 0\}$
- Kolmogorov-Smirnov: simple H_0 & adjusted (composite H_0)
- Chi-square test: simple H_0

different test statistics, with different distributions under H_0

to finish

To summarize

Today we discussed

- Quantile function
- Location-scale family
- QQ-plots and symplots
- Goodness-of-fit tests

Next week goodness-of-fit tests and kernel density estimators