Statistical Data Analysis, Lecture 3

dr. Dennis Dobler

Vrije Universiteit Amsterdam

17 February 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

Chapter 4: Density estimation

Contents of Chapter 4:

1. Kernel density estimators
2. Choice of kernel and bandwidth
3. Cross-validation
4. Other density estimators
5. Multivariate density estimation

Goodness-of-fit (GoF) tests

goodness-of-fit tests

recap hypothesis tests: clearly state

- 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 $x_1, \ldots, x_n$ 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.

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

Shapiro-Wilk test

Shapiro-Wilk test

for composite $H_0 : F \in \{N(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma^2 > 0\}$.
Test statistic:

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \in (0, 1]$$

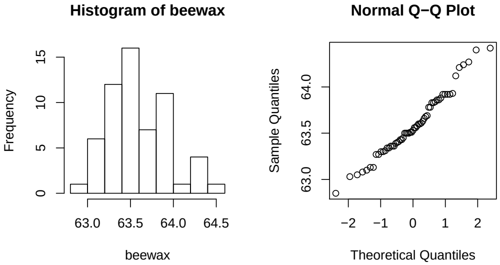
with $a_1, \ldots, a_n$ constants.

Distribution of W under H_0 is known from tables (or R).
Reject H_0 for "small" values of W.

R: shapiro.test

Example Shapiro-Wilk test (1)

Example Beewax data: melting points (°C) of 59 samples of beewax.



Is normality an adequate assumption?

Example Shapiro-Wilk test (2)

Apply test to beewax data:

```
> shapiro.test(beewax)

        Shapiro-Wilk normality test

data:  beewax
W = 0.9748, p-value = 0.2579
```

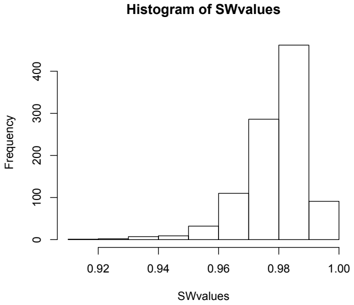
Apply to exponential sample:

```
> shapiro.test(rexp(50))

        Shapiro-Wilk normality test

data:  rexp(50)
W = 0.9026, p-value = 0.0005874
```

Null Distribution of Shapiro-Wilk test statistic

Simulate realizations of it & plot histogram:

```
> SWvalues=numeric(1000)
> for (i in 1:1000)
+ {
+  x=rnorm(59)
+  SWvalues[i]=shapiro.test(x)[[1]]
+ }
> hist(SWvalues)
```

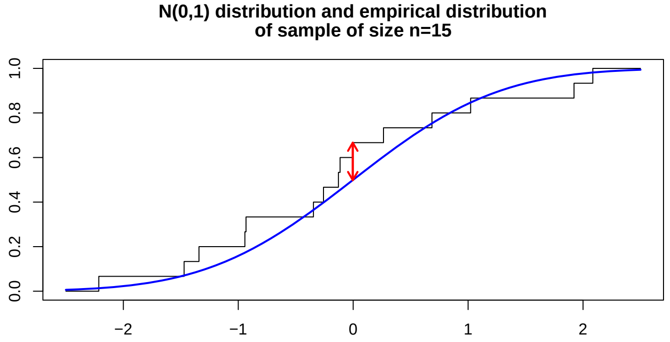


Kolmogorov-Smirnov test

Kolmogorov-Smirnov test (1)

for simple $H_0 : F = F_0$ versus $H_1 : F \neq F_0$.

Test statistic: maximum vertical distance between $\widehat{F}_n$ & F_0:



Kolmogorov-Smirnov test (2)

Test statistic: $D_n = \sup_{-\infty<x<\infty} \left|\widehat{F}_n(x) - F_0(x)\right|.$

H_0 is rejected for large values of D_n.

Null distribution of D_n depends on n, but independent of F_0 if F_0 cont.!

$$D_n = \max_{1\leq i\leq n} \max\left\{\left|\frac{i}{n} - F_0(X_{(i)})\right|, \left|\frac{i-1}{n} - F_0(X_{(i)})\right|\right\}.$$

$\Rightarrow$ KS-test is nonparametric, or distribution free over the class of continuous functions.

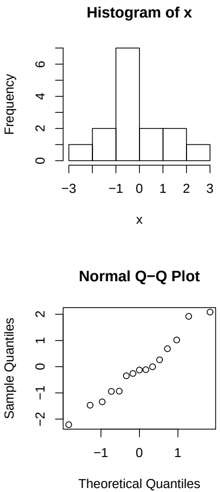
R: ks.test

Example Kolmogorov-Smirnov test

Test $H_0 : X_1, \ldots, X_n \sim N(0, 1)$

```
> ks.test(x,pnorm,0,1)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1681, p-value = 0.73
alternative hypothesis: two-sided
```

H_0 not rejected.



How not to use KS-test (1)

Not for testing composite H_0 of normality (i.e. the complete LSF)!

Wrong KS-test application:

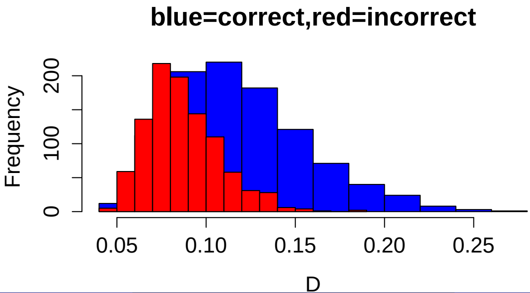
```
> ks.test(x,pnorm,mean(x),sd(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1287, p-value = 0.9378
alternative hypothesis: two-sided
```

Later: bootstrap KS-test version for testing composite normality.

How not to use the Kolmogorov-Smirnov test (2)

Correctly and **incorrectly** computed D_n-values:

```
> dval=numeric(1000)
> tval=numeric(1000)
> for(i in 1:1000) {
+     x=rnorm(50)
+     dval[i]=ks.test(x,pnorm,0,1)[[1]]
+     tval[i]=ks.test(x,pnorm,mean(x),sd(x))[[1]]
+ }
> hist(dval,col="blue",main="blue=correct,red=incorrect",xlab="D")
> hist(tval,add=TRUE,col="red")
```



χ^2 test

Chi-square GoF test (1)

for simple $H_0 : F = F_0$ versus $H_1 : F \neq F_0$.

Test statistic: difference observed – expected number of observations in intervals $I_1, \ldots, I_k$.

Chi-square GoF test (2)

Test statistic (sample of size n):

$$X^2 = \sum_{i=1}^{k} \frac{\left[N_i - np_i\right]^2}{np_i},$$

N_i = observed number of measurements in I_i,
np_i = expected number of measurements in I_i $\quad (p_i = F_0\{I_i\})$.

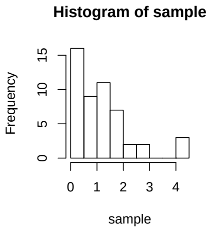
Reject H_0 for large values of X^2.

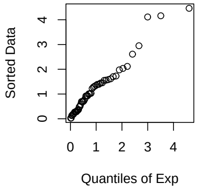
Null distribution of X^2: asymptotically χ^2_{k-1}. (Reliable if all $np_i \geq 5$.)
Also: (asymptotically) distribution free!

R: chisquare (on Canvas)

Example chi-square test (1)

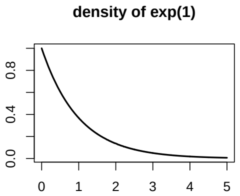
```
> range(sample)
[1] 0.02910324 4.46345348
> length(sample)
[1] 50
> chisquare(sample, pexp, 10, 0, 5)
$chisquare
[1] 26.30088
$pr
[1] 0.001823704
$N
(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5]
     16       9      11       7       2
(2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5]
      2       0       0       3       0
$np
[1] 19 11 7 4 2 1 0 0 0 0
```

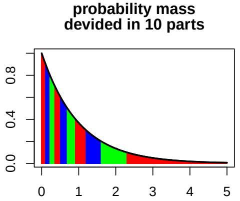




Example chi-square test (2)

```
> b
[1] 0.0 0.1 0.2 0.4 0.5 0.7 0.9 1.2 1.6 2.3  Inf
> chisquare(sample, pexp, 10, 0, 5,b)
$chisquare
[1] 13.6
$pr
[1] 0.1372824
$N
    (0,0.105] (0.105,0.223] (0.223,0.357]
            2             5             6
(0.357,0.511] (0.511,0.693] (0.693,0.916]
            3             1             5
  (0.916,1.2]    (1.2,1.61]    (1.61,2.3]
            6            11             6
    (2.3,Inf]
            5
$np
 [1] 5 5 5 5 5 5 5 5 5 5
```

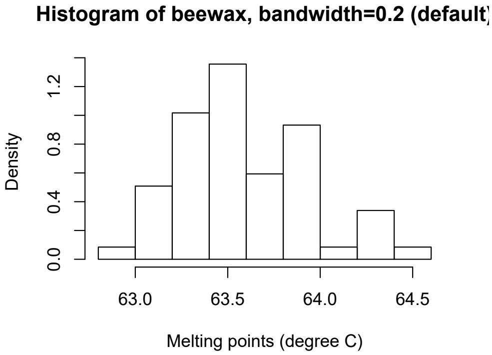




Kernel density estimation

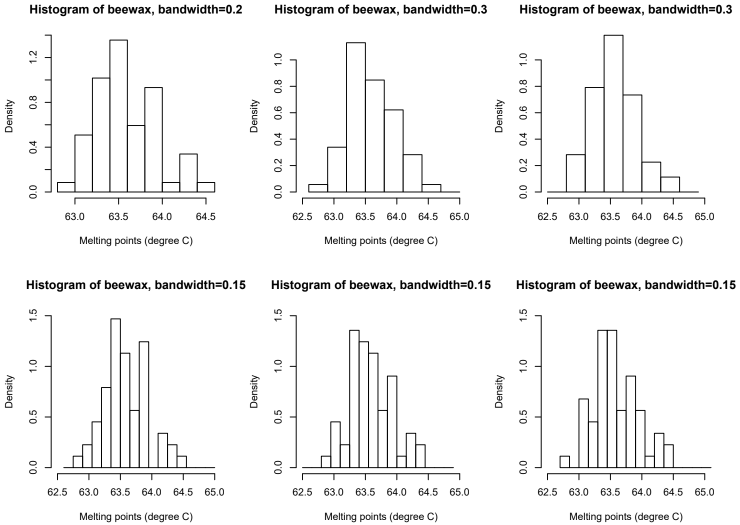
Kernel density estimation (1)

Recall histogram (rescaled to density):
Example Beewax data: melting points (°C) of 59 samples of beewax.



No one believes in such a density.
Other breaks (location/width)?

Kernel density estimation (2)



Kernel density estimation (3)

Aim: nonparametric density estimation with reasonable function-valued estimate.

Kernel density estimator $\hat{f}$: certain estimator of density f.

Let $x_1, \ldots, x_n$ originate from continuous distribution; unknown density f.

$\hat{f}$ distributes mass $\frac{1}{n}$ smoothly around each x_i, according to kernel function K. Bandwidth parameter $h > 0$ specifies spread of mass.

Kernel density estimation (4)

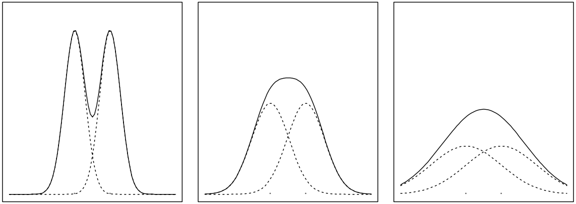
Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$; $P(a < X_1 < b) = \int_a^b f(t)dt$, $-\infty < a < b < \infty$.
K: density function; expectation 0; variance 1. $h > 0$

Kernel density estimator:

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{t - X_i}{h}\Big),$$

kernel K, bandwidth h.
$\hat{f}$ smooth $\Leftrightarrow$ K smooth.

$\hat{f}$ with Gaussian kernel, two observations, different bandwidths:



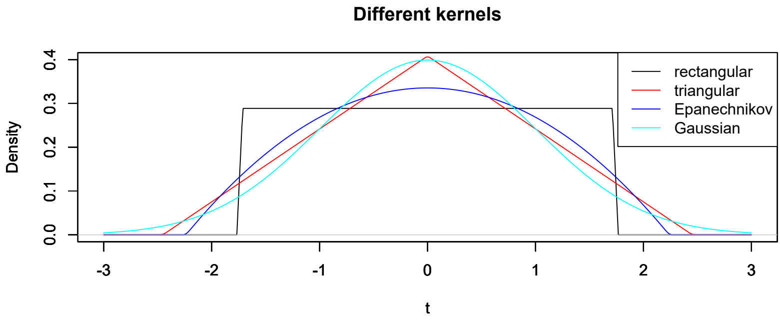
Choice of kernel and bandwidth

Choice of kernel and bandwidth (1)

Problem: how to choose K and h?

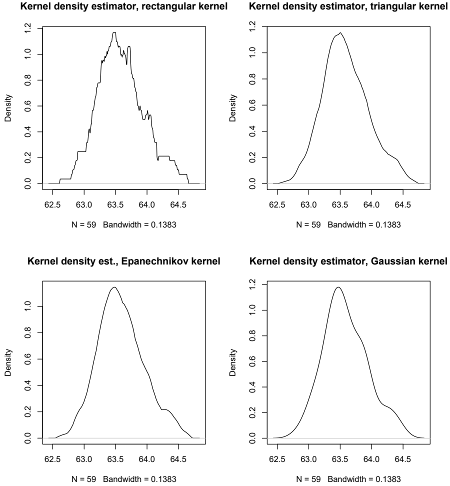
(K less important than h.)

Selection of possible kernel functions:

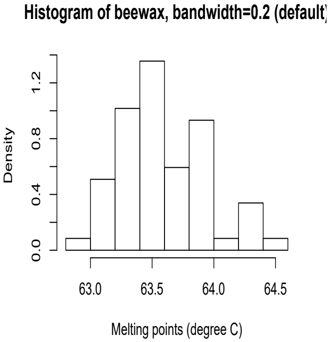


Choice of kernel and bandwidth (2)

Beewax data, different kernels, R's default bandwidth:

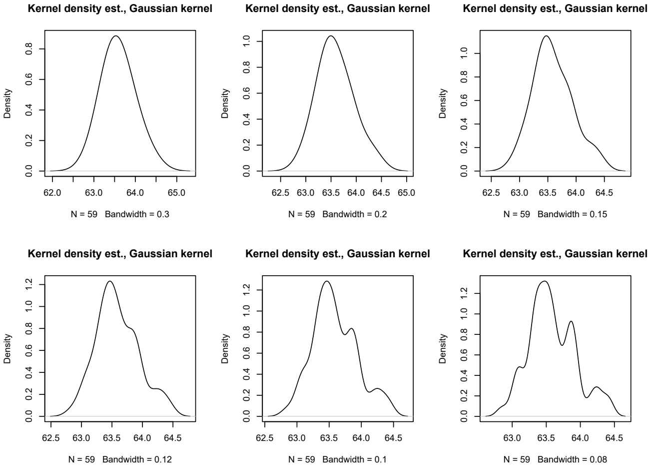


Compare to histogram:



Choice of kernel and bandwidth (3)

As for histograms, bandwidth plays crucial role:




```
plot(density(beewax, bw=0.3, kernel = "gaussian"), ...)
```

to finish

To summarize

Today we discussed

- Goodness-of-fit tests
  - Shapiro-Wilk test
  - Kolmogorov-Smirnov test
  - Chi-square test
- Kernel density estimation
- Choice of kernel and bandwidth

Next week

- Choice of kernel and bandwidth (continued)
- Cross-validation
- Other density estimators
- Multivariate density estimators
- Bootstrap methods