Additional Prep for Test 3: Chapter 13

SS 2141A

Ričardas Zitikis

School of Mathematical and Statistical Sciences
Western University, Ontario

1 Basics

2 Density

3 Problems

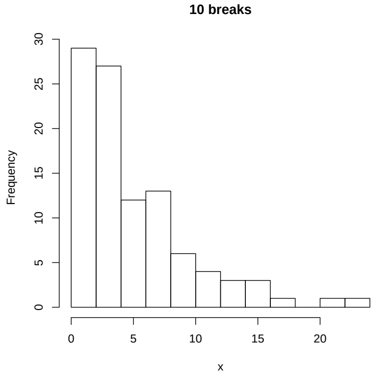
4 Exercises

A toy dataset “x”

3.5, 4.8, 7.3, 1.7, 0.5, 11.6, 6.3, 9.0, 20.9, 8.6, 1.1, 7.8, 5.7, 1.2, 22.6, 3.1, 1.5, 4.4, 2.0, 0.9, 3.6, 3.4, 15.3, 13.0, 2.7, 1.0, 2.4, 10.6, 7.4, 2.9, 7.5, 0.3, 3.9, 9.9, 6.6, 2.2, 15.4, 0.9, 0.9, 9.4, 1.1, 1.5, 3.3, 0.8, 3.2, 3.5, 5.2, 6.5, 1.8, 1.3, 15.0, 7.5, 6.5, 0.8, 5.7, 1.7, 11.5, 2.6, 3.4, 1.7, 12.6, 0.6, 1.5, 4.2, 2.5, 9.2, 0.2, 5.9, 0.5, 0.2, 2.8, 6.0, 10.9, 0.2, 6.7, 4.9, 3.6, 6.1, 2.8, 0.9, 3.1, 3.2, 4.3, 7.6, 2.3, 3.9, 5.6, 2.1, 17.1, 9.8, 4.5, 0.6, 3.7, 0.9, 0.4, 13.3, 2.7, 2.4, 3.3, 6.9

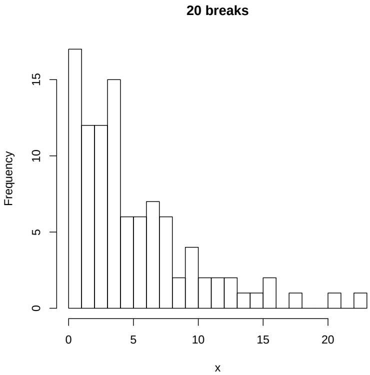
```
> length(x)
[1] 100
> min(x)
[1] 0.2
> max(x)
[1] 22.6
> quantile(x, c(0.25, 0.50, 0.75))
25% 50% 75%
1.7 3.5 7.0
> mean(x)
[1] 5.084
> var(x)
[1] 21.44196
> sqrt(var(x))
[1] 4.630547
```

```
> hist(x,breaks=10,main="10␣breaks")
```



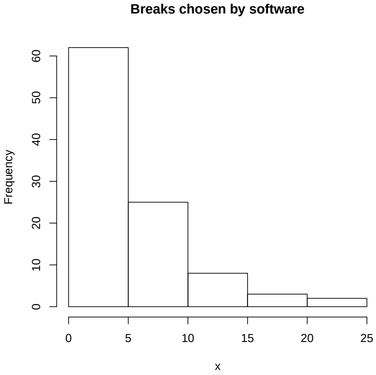
I asked the computer for 10 bins, but it decided that the choice was poor and opted for 12 bins instead. Let's not argue with it.

```
> hist(x,breaks=20,main="20 breaks")
```



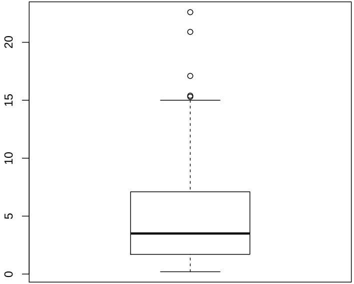
I asked for 20 bins but the computer seems to have adjusted my choice again. Let it be so.

```
> hist(x,main="Breaks␣chosen␣by␣software")
```



Here is the computer's own choice of bins.

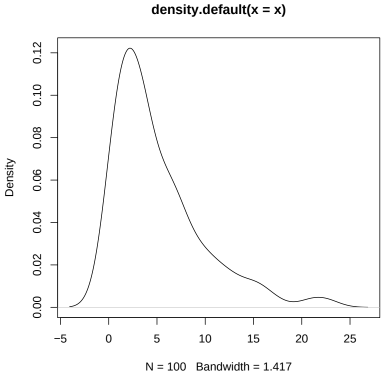
```
> boxplot(x)
```



The computer thinks that some data points are outliers and indicates them using circles. This is the so-called modified boxplot.

1. Basics
2. Density
3. Problems
4. Exercises

```
> plot(density(x))
```



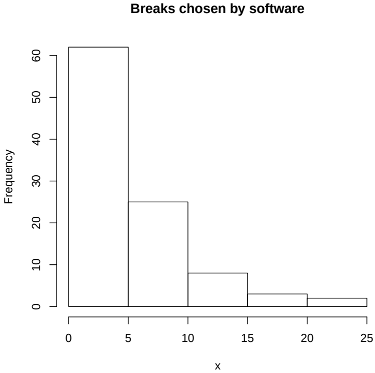
The computer did its best to fit a density curve. It went even into negative numbers, although the dataset does not have such values.

- Let's denote the density function by $f(x)$.
- If we want to know the proportion of the data points between some numbers a and b, then we can (approximately) find this proportion by calculating the area underneath the density function between a and b.
- This area is nothing else but the integral

$$\int_a^b f(x)\mathrm{d}x$$

- Let's choose $a = 5$ and 15. To calculate the integral $\int_5^{15} f(x)\mathrm{d}x$, we need a formula for $f(x)$, but the computer just fits a function but does not give a formula. What to do?

Let's look at the computer's drawn histogram



The decay looks exponential, and so let's fit $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

The exponential function is $e^{-\lambda x}$, but why do we multiply it by λ to get $f(x) = \lambda e^{-\lambda x}$?

The reason is that the area under the density function (just like the area of every frequency histogram) has to be equal to 1.

Let's check:

$$\begin{aligned}\int_0^\infty f(x)\mathrm{d}x &= \int_0^\infty \lambda e^{-\lambda x}\mathrm{d}x \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}x}\Big(-e^{-\lambda x}\Big)\mathrm{d}x \\ &= -e^{-\lambda x}\Big|_0^\infty \\ &= 1\end{aligned}$$

So what is the (approximate) proportion of the data points between $a = 5$ and 15?

Let's calculate

$$\begin{aligned}\int_5^{15} f(x)\mathrm{d}x &= \int_5^{15} \lambda e^{-\lambda x}\mathrm{d}x \\ &= \int_5^{15} \frac{\mathrm{d}}{\mathrm{d}x}\Big(- e^{-\lambda x}\Big)\mathrm{d}x \\ &= -e^{-\lambda x}\Big|_5^{15} \\ &= -\Big(e^{-\lambda 15} - e^{-\lambda 5}\Big) \\ &= e^{-\lambda 5} - e^{-\lambda 15}\end{aligned}$$

What is λ? If we know it, we know the numerical value of the integral (i.e., of the proportion) $\int_5^{15} f(x)\mathrm{d}x$ using the above formula.

It appears that λ is the reciprocal of the mean.

Recall from our earlier calculations that the mean is

$$\bar{x} = 5.084$$

and so

$$\lambda = \frac{1}{5.084} = 0.1966955$$

Hence,

$$\begin{aligned}\int_5^{15} f(x)\mathrm{d}x &= e^{-\lambda 5} - e^{-\lambda 15} \\ &= e^{-0.1966955\times 5} - e^{-0.1966955\times 15} \\ &= 0.3216911\end{aligned}$$

Using the exponential density, we have arrived at the value

$$\int_{5}^{15} f(x)\mathrm{d}x = 0.3216911$$

But what is the true proportion of data points between 5 and 15?

Let's run a little code:

```
> x5to15 <- x[5 <= x & x <= 15]
> length(x5to15)/length(x)
[1] 0.33
```

Hence, the true proportion of x's between 5 and 15 is

$$0.33$$

Compare it with 0.3216911. The two are quite close, aren't they?

1 Basics

2 Density

3 Problems

4 Exercises

How to find the median?

- Suppose that instead of a histogram, we are now given a density function. How can we find the median?
- Recall that the median splits the data in half, which means that the median (approximately) splits the histogram's area in half.
- Since we now have a density function, this means that the median splits the area under the density function in half.
- Hence, the median M is the solution to the equation

$$\int_M^\infty f(x)\mathrm{d}x = 0.5$$

What is M?

Suppose that we work with the exponential density function

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Hence

$$\begin{aligned}
\int_M^\infty f(x)\mathrm{d}x &= \int_M^\infty \lambda e^{-\lambda x}\mathrm{d}x \\
&= \int_M^\infty \frac{\mathrm{d}}{\mathrm{d}x}\Big(- e^{-\lambda x}\Big)\mathrm{d}x \\
&= -e^{-\lambda x}\Big|_M^\infty \\
&= -\Big(0 - e^{-\lambda M}\Big) \\
&= e^{-\lambda M}
\end{aligned}$$

Consequently, to find the median M, we need to solve the equation

$$e^{-\lambda M} = 0.5$$

It is equivalent to

$$-\lambda M = \log(0.5)$$

and thus

$$\begin{aligned} M =& \frac{-\log(0.5)}{\lambda} \\ &= \frac{\log(2)}{\lambda} \end{aligned}$$

because $-\log(1/2) = \log(2)$.

How to find the mean?

- Suppose again that instead of a histogram, we are given a density function. How can we find the mean?
- The mean of a density function, which is usually denoted by the Greek letter μ, is given by the formula

$$\mu = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$$

- To illustrate, suppose that we work with the exponential density function

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Since the exponential density function $f(x)$ is defined for only $x \geq 0$, meaning that $f(x) = 0$ for all $x < 0$, the mean is

$$\begin{aligned}\mu = \int_0^\infty x f(x) \mathrm{d}x &= \int_0^\infty x \lambda e^{-\lambda x} \mathrm{d}x \\ &= \int_0^\infty x \frac{\mathrm{d}}{\mathrm{d}x}\Big(- e^{-\lambda x}\Big) \mathrm{d}x \\ &= -x e^{-\lambda x}\Big|_0^\infty - \int_0^\infty \Big(- e^{-\lambda x}\Big) \mathrm{d}x \\ &= 0 + \int_0^\infty e^{-\lambda x} \mathrm{d}x \\ &= \frac{1}{\lambda}\end{aligned}$$

Compare the mean with the median $M = \log(2)/\lambda$. Which one, μ or M, is larger?

1 Basics

2 Density

3 Problems

4 Exercises

Exercise: find the standard deviation

- Again, we are given a density function.
- Its standard deviation, usually denoted by the Greek letter σ, is the square root of the variance σ^2. (A bit of mathematical tautology.)
- The variance is given by the formula

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \mathrm{d}x$$

- Suppose that we work with the exponential density function

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Calculate its standard deviation σ. (Recall that $\mu = 1/\lambda$.)

Exercise: exponential five-number summary

We know that the median of the exponential distribution is

$$M = \frac{\log(2)}{\lambda}.$$

It is also called the 2nd quartile Q_2, or the 50th percentile.

Complete the five-number summary for the exponential distribution, that is, find:

- the minimum
- the 1st quartile Q_1, also known as the 25th percentile
- the 3rd quartile Q_3, also known as the 75th percentile
- the maximum

Exercise: Weibull five-number summary

Like the exponential, the Weibull distribution is very popular in engineering research and practice (see, e.g., weibull.com). Its density function $f(t)$ satisfies the equation (i.e., the area to the right of x)

$$\int_x^\infty f(t)\mathrm{d}t = e^{-\lambda x^\alpha}, \quad x \geq 0,$$

with some parameters $\lambda > 0$ and $\alpha > 0$. Find the five-number summary:

- the minimum
- the 1st quartile Q_1, also known as the 25th percentile
- the 2nd quartile Q_2, also known as the median M or the 50th percentile
- the 3rd quartile Q_3, also known as the 75th percentile
- the maximum