Histogram

Show the data distribution and suggests possible outliers. Its shape is similar to the population pdf/pmf, especially if the sample size is large.

Frequency histogram: Consists of bars, one over each bin, whose heights represent the *number* of observations in the bins.

Relative frequency histogram: Consists of bars, one over each bin, whose heights represent the *proportion* of observations in the bins.

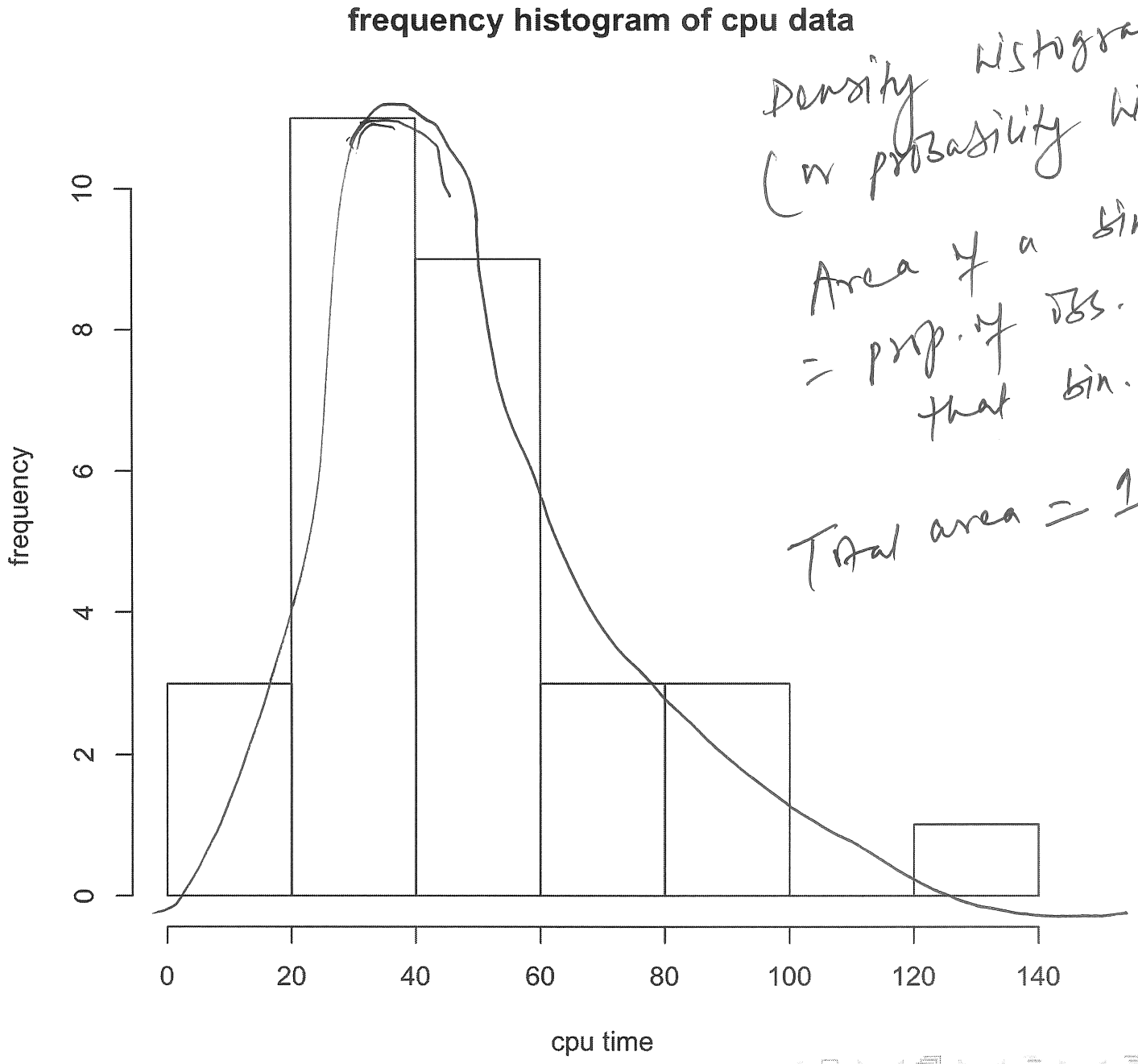
How to construct a histogram?

- effect of number of bins (too many or too few)
- bins of unequal sizes

Don't change default in R unless there is a good reason.

```
# frequency histogram by default
hist(cpu, xlab="cpu time", ylab="frequency",
     main="histogram of cpu data")

# probability (density) histogram
hist(cpu, freq=FALSE, xlab="cpu time",
ylab="density", main="histogram of cpu data")
```

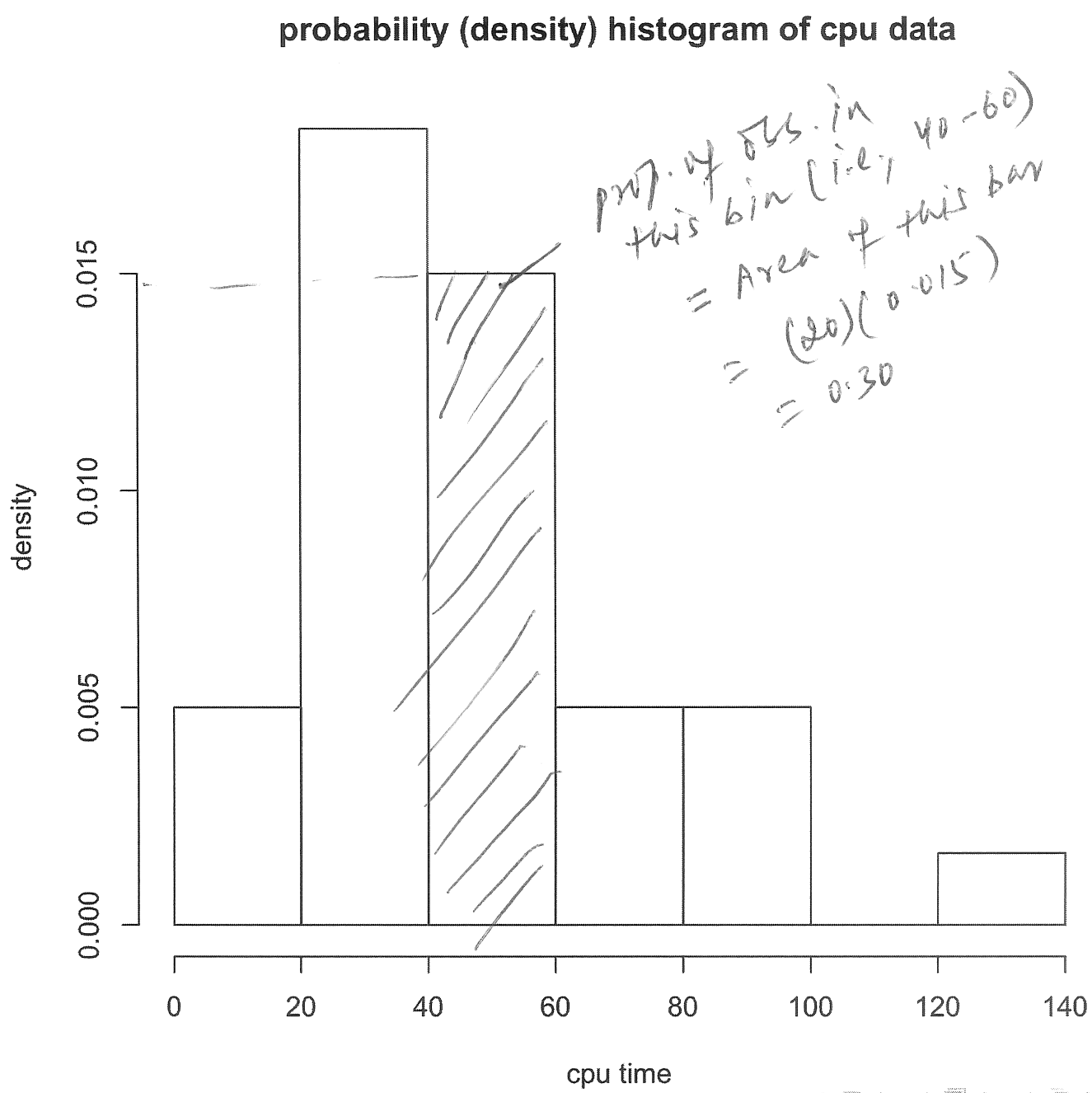
Find a way to make a rel. freq. hist. in R.



Density histogram
(or probability histogram):

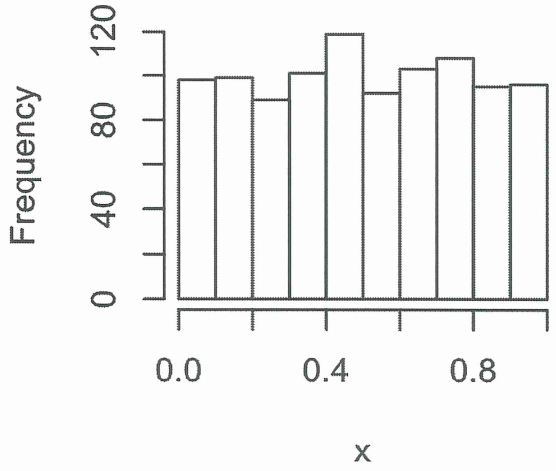
Area of a bin
= prop. of obs. of in
that bin.

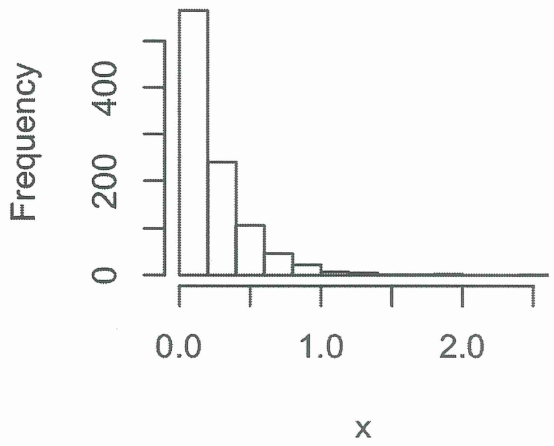
Total area = 1.

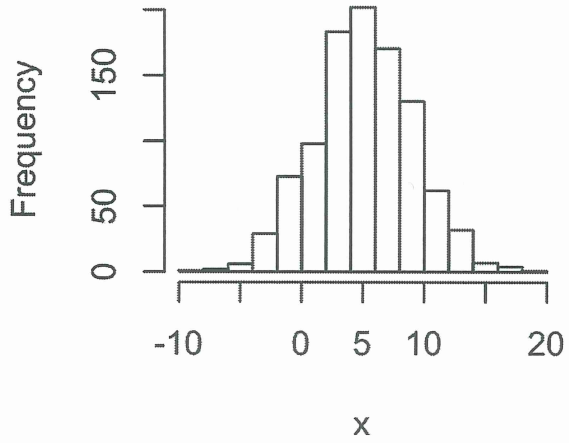

probability (density) histogram of cpu data
density
0.000
0.005
0.010
0.015
0
20
40
60
80
100
120
140
cpu time
prob. of obs. in this bin (i.e., 40-60)
= Area of this bar
= (20)(0.015)
= 0.30

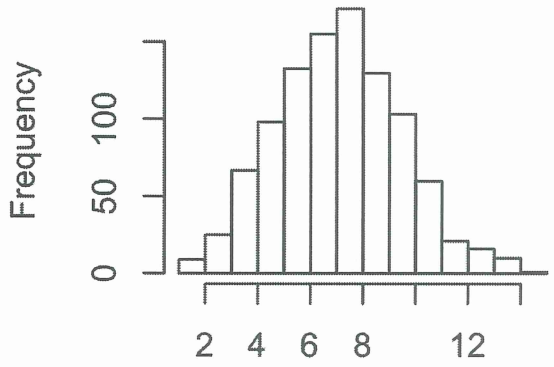
Histograms of some simulated data:

```
nsim <- 1000
# uniform (0,1) distribution
par(mfrow=c(2,2))
hist(runif(nsim), xlab="x", main="")

# exponential (lambda = 4) distribution
hist(rexp(nsim, rate=4), xlab="x", main="")

# normal (mu=5, sigma^2=16) distribution
hist(rnorm(nsim, mean=5, sd=4), xlab="x", main="")

# binomial (n=30, p=0.25)
hist(rbinom(nsim, size=30, prob=0.25), main="")
par(mfrow=c(1,1))
```









rbinom(n=sim, size = 20, prob = 0.25)

Why does the last histogram have a "normal shape?"

This is because of CLT:

$$\text{Bin}(n, p) \approx \text{Normal}\left(\mu = np,\ \sigma^2 = np(1-p)\right),$$

provided n is large.

Rule of thumb: $np \geq 5$ and $np(1-p) \geq 5$

QQ Plot

Plot quantiles of one dataset against quantiles of another dataset (from a known distribution with cdf F). If the points fall on a straight line, the distribution F may be a good fit to the data — allows a graphical check of how well F fits the data.

Data: $x_1, \ldots, x_n$ (a random sample)

Sorted data: $x_{(1)}, \ldots, x_{(n)}$

- These are sample quantiles or "order statistics."
- They estimate population quantiles of the distribution F (assuming that F is the correct model for the data)

Q: What are the associated probabilities?

- Each sample observation has $1/n$ probability weight under the empirical distribution.
- The sample quantiles $x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)}$ are associated with probabilities $1/n, 2/n, \ldots, (n-1)/n, n/n$.
- $x_{(i)}$ estimates (i/n)th population quantile, i.e., $F^{-1}(i/n)$, $i = 1, \ldots, n$.

p	$x_{(np)}$	$F^{-1}(p)$
$\frac{1}{n}$	$x_{(1)}$	$F^{-1}(p=\frac{1}{n})$
$\frac{2}{n}$	$x_{(2)}$	$F^{-1}(p=2/n)$
$\frac{n-1}{n}$	$x_{(n-1)}$	$F^{-1}(p=\frac{n-1}{n})$
n	$x_{(n)}$	$F^{-1}(p=1)$

↑ sample quantile with $p=\frac{1}{n}$

$p=2/n$

$p=\frac{n-1}{n}$

$p=\frac{n}{n}=1$

↑ max. in the pop.

QQ plot: Plot the following pairs of points: $(x_{(i)}, F^{-1}(i/n))$, $i = 1, \ldots, n$.

Problem: $F^{-1}(1)$ may be ∞.

Solution: Consider an offset a.

Old probabilities: $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}$

New probabilities: $\frac{1-a}{n+1-2a}, \frac{2-a}{n+1-2a}, \ldots, \frac{n-1-a}{n+1-2a}, \frac{n-a}{n+1-2a}$

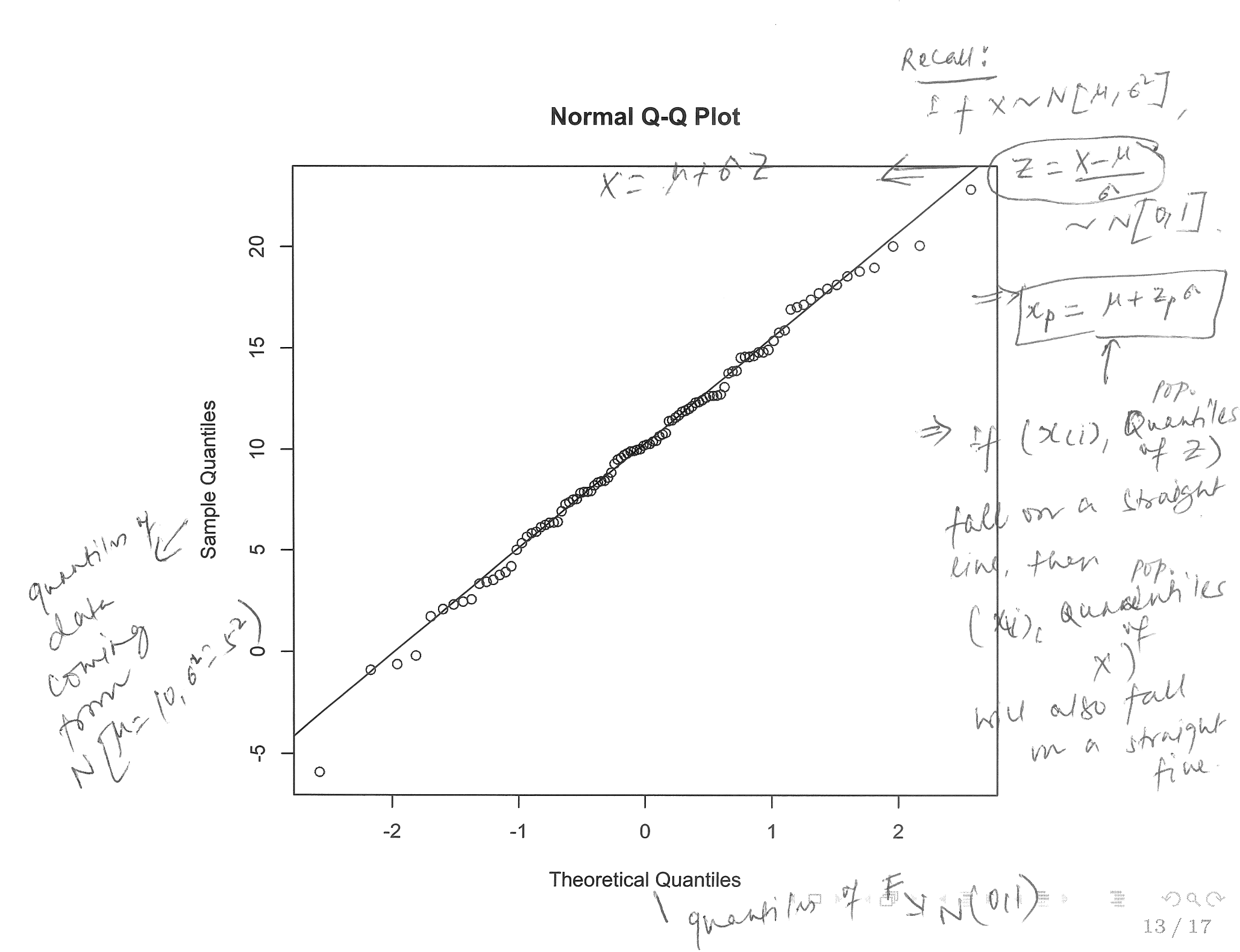
- Default in R: $a = 3/8$ if $n \leq 10$ and $a = 1/2$ if $n > 10$.
- `qqplot` gives a general QQ plot
- `qqnorm` gives normal QQ plot — it uses $N(0,1)$ distribution as F

Q: What are the probabilities for $n = 30$?

(with offset).

$n = 30 \hookrightarrow a = \frac{1}{2}$

$\frac{1-\frac{1}{2}}{(30+1-1)} = \frac{0.5}{30}, \frac{1.5}{30}, \ldots, \frac{29.5}{30}$



Recall:

If $X \sim N[\mu, \sigma^2]$,

$Z = \frac{X-\mu}{\sigma} \sim N[0,1]$.

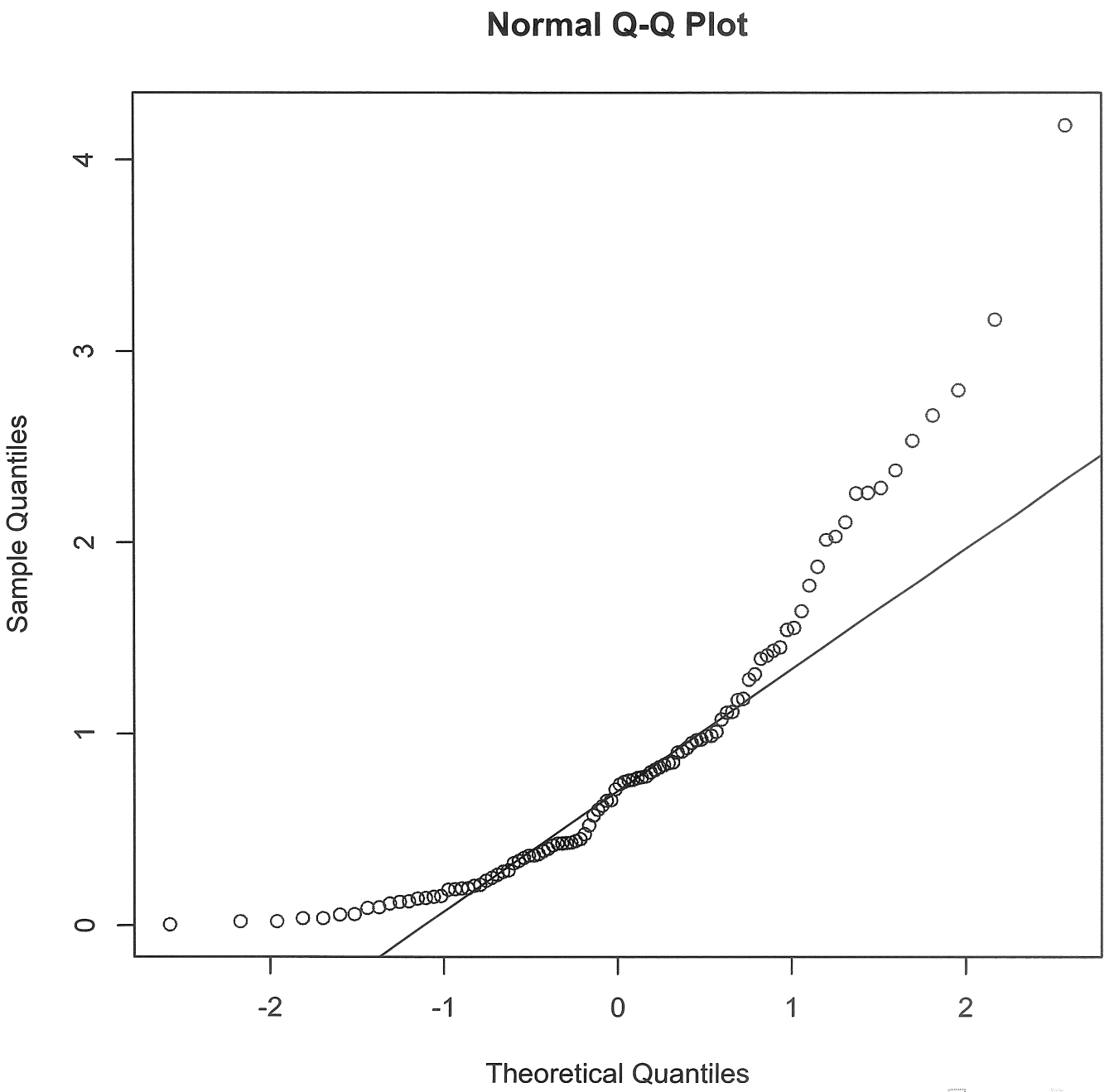
$X = \mu + \sigma Z$

$x_p = \mu + z_p \sigma$

$\Rightarrow$ If $(x_{(i)}$, pop. Quantiles of $Z)$ fall on a straight line, then pop. $(x_{(i)}$, Quantiles of $X)$ will also fall on a straight line.

quantiles of data coming from $N[\mu=10, \sigma^2=5^2]$

quantiles of $F_{N(0,1)}$


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
-2
-1
0
1
2

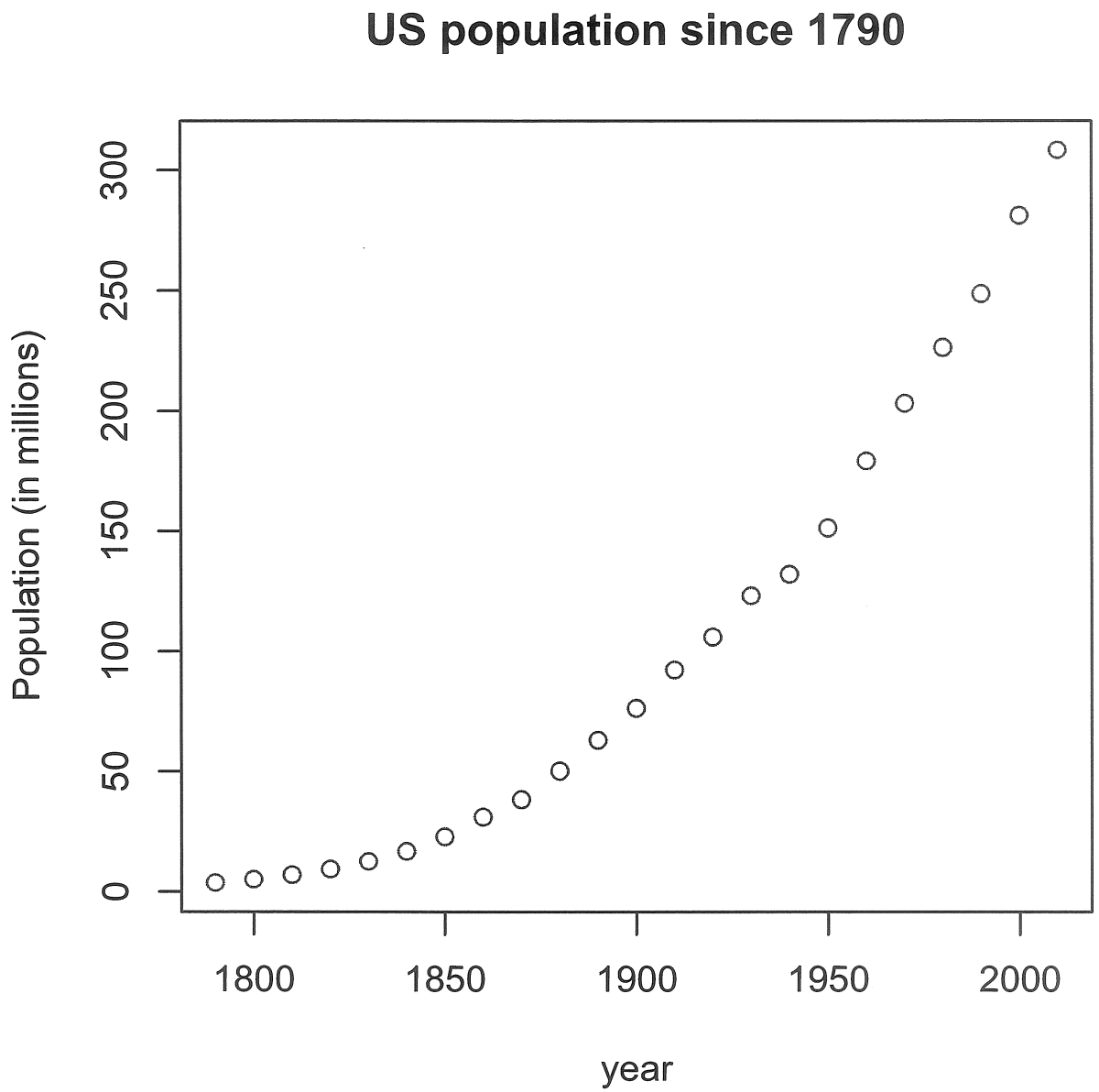
R Code for QQ plots

```
# QQ plot 1
x <- rnorm(100, 10, 5)
qqnorm(x)
qqline(x)

# QQ plot 2
x <- rexp(100, 1)
qqnorm(x)
qqline(x)
```

Time series plot: Plot of a data on a variable against time — shows how the variable changes over time.

```
# Data from Exercise 8.5
year <- seq(from=1790, to=2010, by=10)
# > year
 # [1] 1790 1800 1810 ...  2010
# >
uspop <- c(3.9, 5.3, 7.2, 9.6, ..., 281.4, 308.7)

plot(year, uspop, ylab="Population (in millions)",
  main="US population since 1790")
```

Scatterplot: Plot of one variable (X) against another variable Y — shows the relationship between the two variables. See Figure 8.11 of the textbook.


US population since 1790
Population (in millions)
0
50
100
150
200
250
300
1800
1850
1900
1950
2000
year

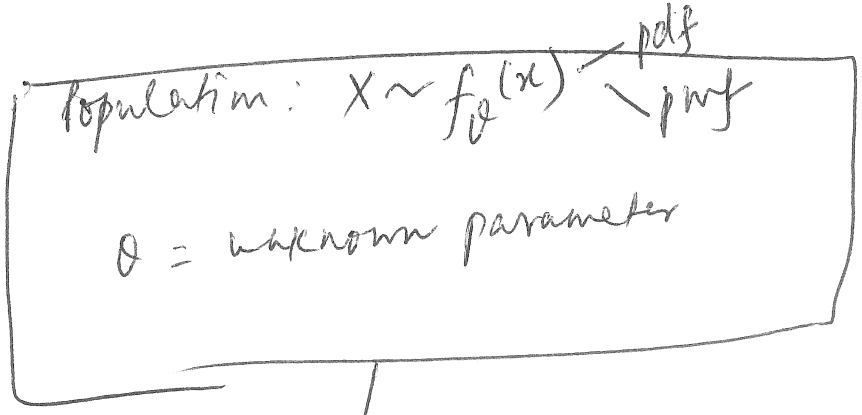
Point estimation (Chapter 9)

Problem: $X \sim f_\theta(x)$, where θ is an unknown parameter. This θ may be a vector.

Data: $X_1, \ldots, X_n$ — a random sample of X.

We have seen a number of descriptive statistics and what they estimate. But the choice of an $\hat{\theta}$ of θ may not be obvious.

A general method of parameter estimation: Method of maximum likelihood. It has generally good properties.



Method of Maximum Likelihood

Likelihood function of data: Joint pdf or pmf of sample data considered as a function of θ with data held fixed at the observed values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

$$L(\theta) = L(\theta; x_1, x_2, , x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

(indep.)

joint pdf/pmf: θ is fixed, think of this a fn. of $x_1, \ldots, x_n$

Likelihood fn: $x_1, \ldots, x_n$ fixed, think of this as a fn. of θ.

- A function of θ — the data are held fixed.

Maximum likelihood estimator (MLE) of θ: The value $\hat{\theta}$ of θ that maximizes the likelihood function as a function of θ.

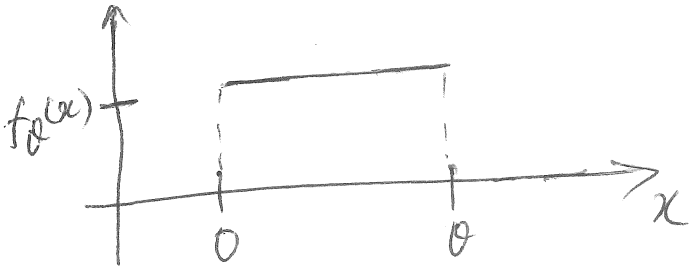
- Can think of MLE as the value of θ that is "most likely" to have led to the observed data.
- Essentially a calculus problem.

How to find MLE?

Direct approach: Directly maximize the likelihood function.

Ex: Let $X_1, X_2, \ldots, X_n$ represent a random sample from a Uniform $(0, \theta)$ distribution where $\theta > 0$. Find the MLE of θ.

Recall: $f_\theta(x) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{o/w} \end{cases}$



PDF of X, which represents the population

Clearly, θ = max. in the population.

will see that: $\hat{\theta}$ = max. in the sample is MLE of θ.

NOTE: $E[X] = \frac{\theta}{2} \Rightarrow \theta = 2E[X]$ — estimate this by $\bar{X}$

$\Rightarrow \hat{\theta} = 2\bar{X}$ can be used as an estimator also of θ.

Have two estimators of θ:

$\hat{\theta}_{MLE} = X_{(n)}$

$\hat{\theta}_{MOME} = 2\bar{X}$

which one is better?