Graphical Statistics

"Plot the data before you do anything with it."

Boxplot: Displays the 5-number summary of the data, i.e., $(\min, \hat{Q}_1, \hat{Q}_2, \hat{Q}_3, \max)$. It shows

- the data distribution (e.g., symmetric, right-skewed or left-skewed)
- outliers

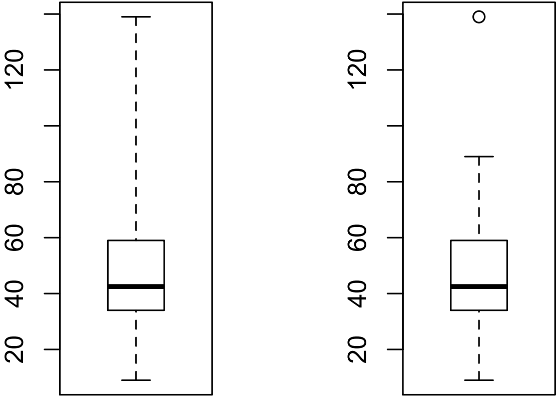
Alternative form: The bottom whisker extends from $\hat{Q}_1$ to $\max\{\min, \hat{Q}_1 - 1.5 \times I\hat{Q}R\}$ and the top whisker extends from $\hat{Q}_3$ to $\min\{\max, \hat{Q}_3 + 1.5 \times I\hat{Q}R\}$

Side-by-side boxplots: Draw side-by-side boxplots on the same scale to compare distributions of more than one data set — see Figure 8.10 in the textbook.

Ex: CPU data

```
?boxplot # see help
par(mfrow=c(1,2)) # 2 plots in 1 row
# plot of 5-number summary
boxplot(cpu, range=0)
# uses 1.5 (IQR) rule (also default), i.e.,
# same as boxplot(cpu)
boxplot(cpu, range=1.5)
par(mfrow=c(1,1)) # back to the default, 1 plot per row
```

Boxplots for CPU data



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.

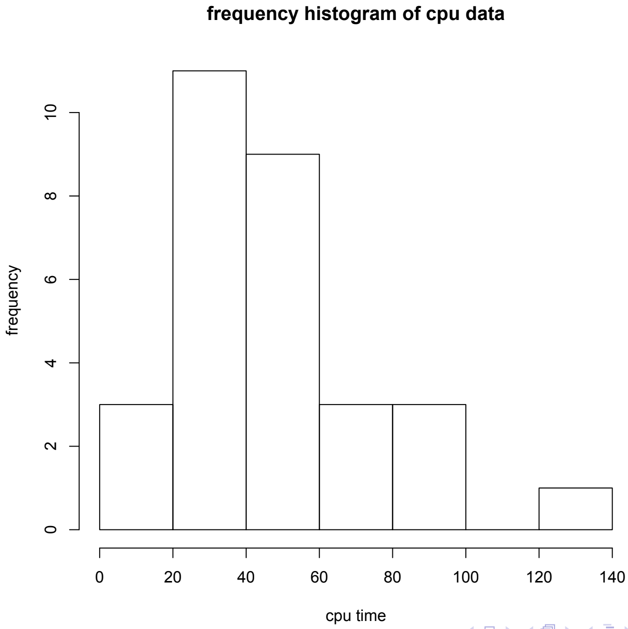
Frequency histogram: Consists of bars, one over each bin, whose heights represent the *number* of observations in the bins.

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

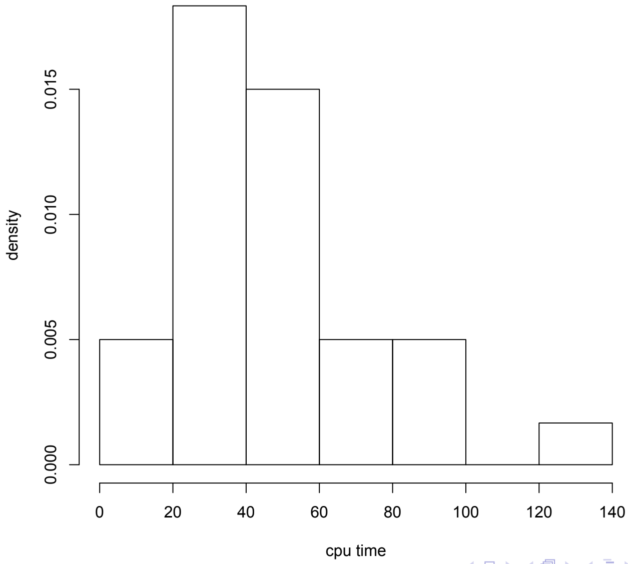
How to construct a histogram?

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

```
# 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")
```

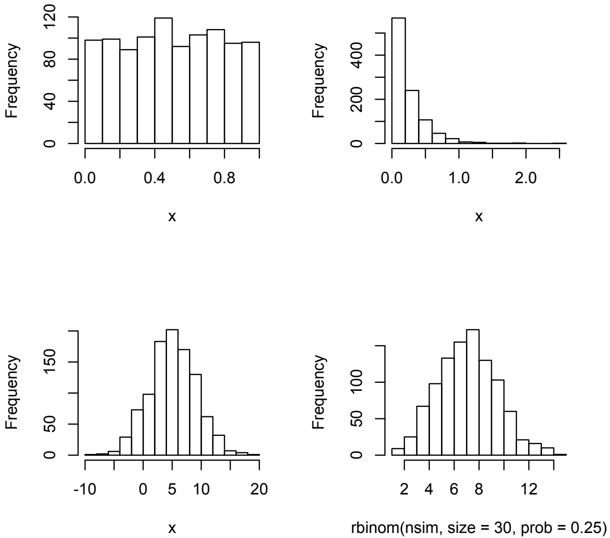

frequency histogram of cpu data
frequency
0
2
4
6
8
10
0
20
40
60
80
100
120
140
cpu time

probability (density) histogram of cpu data



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))
```


Frequency
0
40
80
120
0.0
0.4
0.8
x
Frequency
0
200
400
0.0
1.0
2.0
x
Frequency
0
50
150
-10
0
5
10
20
x
Frequency
0
50
100
2
4
6
8
12
rbinom(nsim, size = 30, prob = 0.25)

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

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.

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., $F^{-1}(i/n)$, $i = 1, \ldots, n$.

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

Problem: $F^{-1}(1)$ may be ∞.

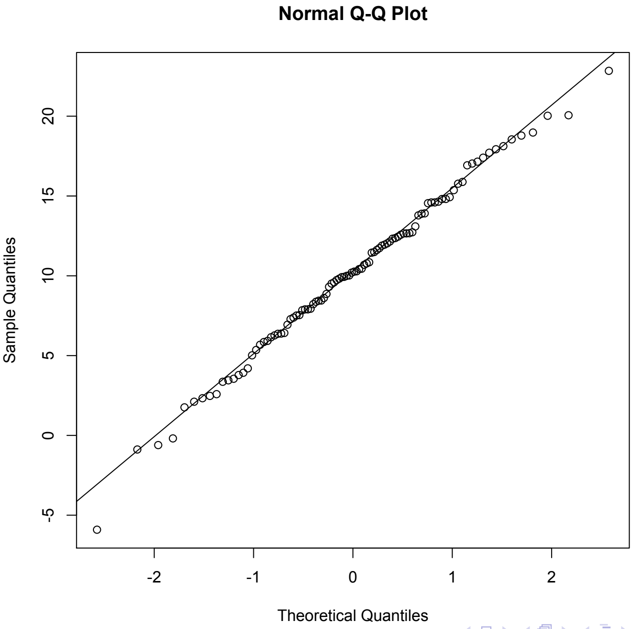
Solution: Consider an offset a.

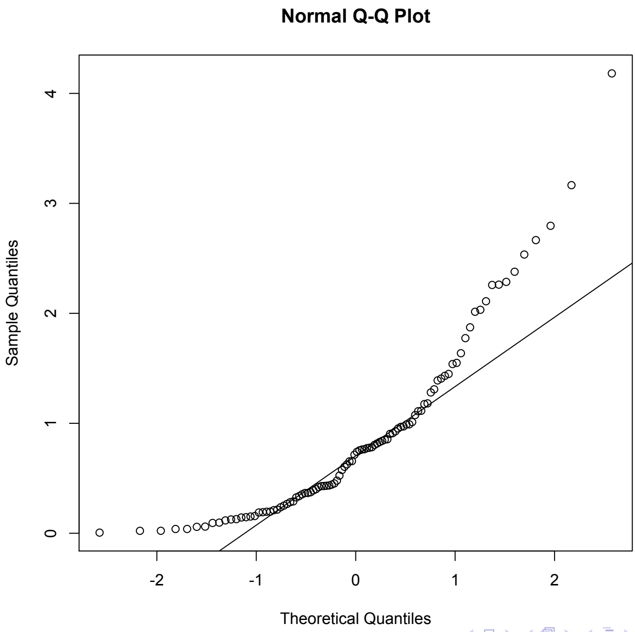
Old probabilities: $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}$

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 $n = 30$?


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-5
0
5
10
15
20
-2
-1
0
1
2


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

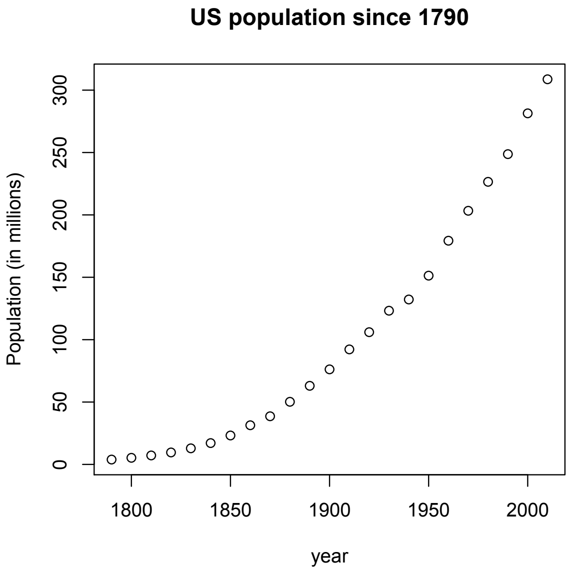
R Code for QQ plots

```
# 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