Data Summaries: Quantiles

```
> library(MASS)
> abbey
 [1]   5.2   6.5   6.9   7.0   7.0   7.0
 [7]   7.4   8.0   8.0   8.0   8.0   8.5
[13]   9.0   9.0  10.0  11.0  11.0  12.0
[19]  12.0  13.7  14.0  14.0  14.0  16.0
[25]  17.0  17.0  18.0  24.0  28.0  34.0
[26] 125.0
> summary(abbey)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.20    8.00   11.00   16.01   15.00  125.00
> min(abbey)
[1] 5.2
> max(abbey)
[1] 125
> range(abbey)
[1]   5.2 125.0
> diff(range(abbey))
[1] 119.8
> median(abbey)
[1] 11
> quantile(abbey, .30)
30%
  8
> quantile(abbey, c(0,.05,.25,.5,.75,.95,1))
   0%    5%   25%   50%   75%   95%  100%
  5.2   6.7   8.0  11.0  15.0  31.0 125.0
> quantile(abbey)
   0%   25%   50%   75%  100%
  5.2   8.0  11.0  15.0 125.0
> IQR(abbey)
[1] 7
>
```


Data Summaries: Univariate Stats

Sample mean	$\bar{x} = \sum_{i=1}^{n} x_i$	`> mean(abbey)` `[1] 16.00645`
Sample variance	$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	`> var(abbey)` `[1] 452.3733`
Sample std. dev.	$s_X = \sqrt{s_X^2}$	`> sd(abbey)` `[1] 21.26907`
Sample covariance	$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$	`> cov(bp, pres)` `[1] 17.34638`
Sample correlation	$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$	`> cor(bp, pres)` `[1] 0.9972102`

Boxplots!!

- Graphical display of Tukey's five number summary: minimum, lower hinge F_L, median, upper hinge F_U, and maximum.

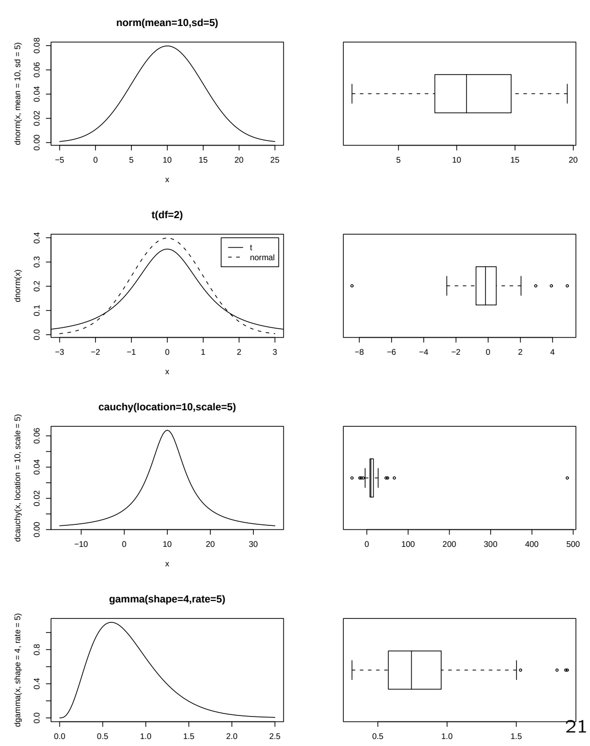
```
> quantile(abbey)
   0%   25%   50%   75%  100%
  5.2   8.0  11.0  15.0 125.0
> fivenum(abbey)
[1]   5.2   8.0  11.0  15.0 125.0
> sample <- 1:4
> quantile(sample)
  0%  25%  50%  75% 100%
1.00 1.75 2.50 3.25 4.00
> fivenum(sample)
[1] 1.0 1.5 2.5 3.5 4.0
>
```

- However, anything outside the range

$$F_L - 1.5\, d_F \quad \text{and} \quad F_U + 1.5\, d_F$$

(where $d_F = F_U - F_L$ is the interhinge range) is considered an outlier and plotted separately.

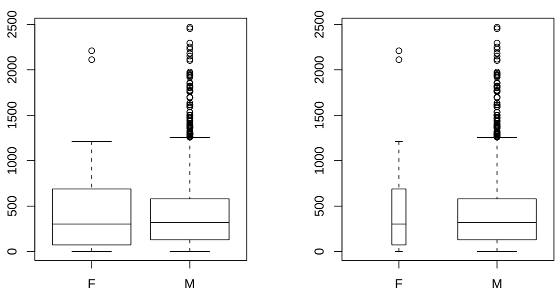
Samples ($n = 50$) from Some Dists



Interesting Variations

- Box widths proportional to $\sqrt{n}$:

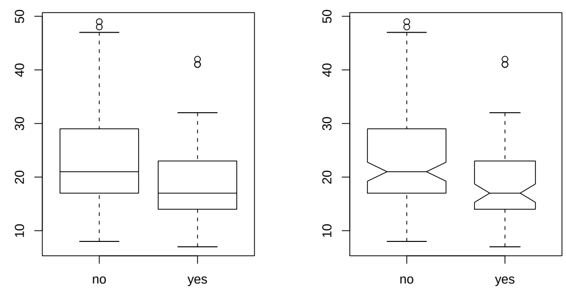
```
> library(MASS)
> attach(Aids2)
> names(Aids2)
[1] "state"   "sex"     "diag"    "death"   "status"
[6] "T.categ" "age"
> boxplot((death-diag) ~ sex)
> boxplot((death-diag) ~ sex, varwidth=T)
> table(sex)
sex
   F    M
  89 2754
>
```



Interesting Variations

- Box notches: no overlap ⇒ very likely difference

```
> library(MASS)
> boxplot(y ~ limit, data=Traffic)
> boxplot(y ~ limit, data=Traffic, varwidth=T, notch=T)
>
```

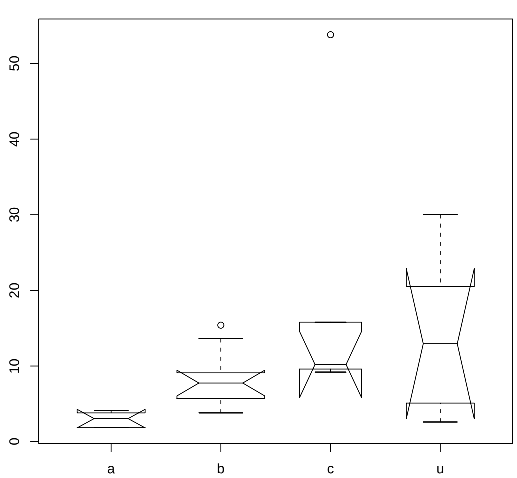


Notches drawn to span *median* $\pm 1.58\, IHR/\sqrt{n}$ (a little bigger than an approximate 90% CI for the median).

Interesting Variations

- Another box-notch and variable width example:

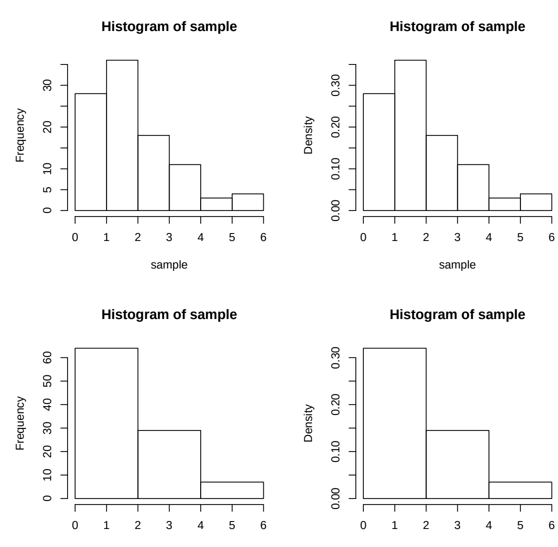
```
> library(MASS)
> boxplot(Tetrahydrocortisone ~ Type, data=Cushings,
          varwidth=T, notch=T)
>
```



Histograms!

. . . by count (`freq=T`) or density (`freq=F`)

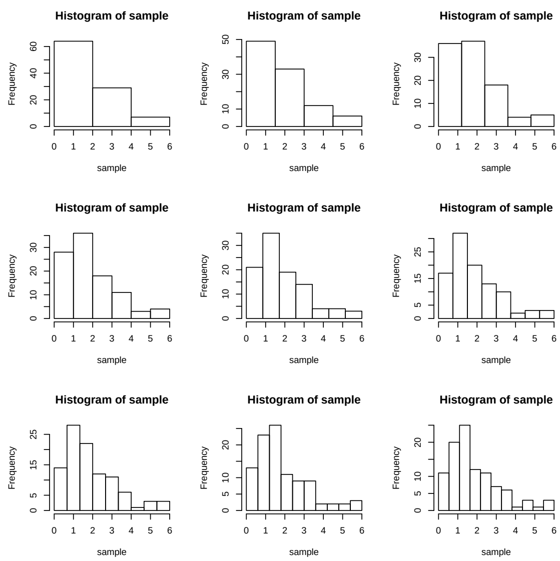
```
> sample <- rgamma(100, shape=2, rate=1)
> hist(sample)          # box height is count
> hist(sample, freq=F)  # box area is proportion of obs
> hist(sample, breaks=3)
> hist(sample, freq=F, breaks=3)
```



Histograms!

. . . are sensitive to choice of breaks.

```
> opar <- par(mfrow=c(3,3))
> for (i in 4:12) hist(sample,breaks=seq(from=0,to=6,length=i))
> par(opar)
```



Histogram Breaks

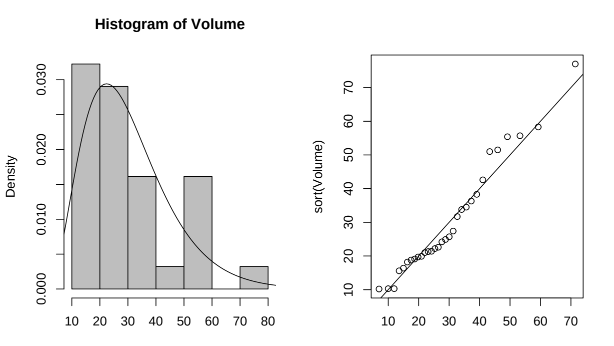
The parameter `breaks` can be given as

- a *suggested* number of breaks `breaks=10`;
- the name of an algorithm to generate a *suggested* number of breaks:
  - `"sturges"`: number of breaks is $\log_2(n) + 1$
  - `"scott"`: box width is $3.5 s n^{-1/3}$
  - `"fd"`: box width is $2\, IQR\, n^{-1/3}$

  This default is `"sturges"` which looks okay for "normalish" data but isn't resistant to outliers.
- a vector of exact breaks (which may be unequally spaced).

Histograms w/ Density Curves

```
> library(MASS)  # needed for fitdistr
> attach(trees)
> hist(Volume, freq=F, col="grey")
> fitdistr(Volume,"gamma")
     shape         rate
  3.88581759   0.12880083
[ . . . ]
> curve(dgamma(x, shape=3.89, rate=0.129),add=T)
> plot(qgamma(ppoints(Volume), shape=3.89, rate=0.129), sort(Volume))
> abline(0,1)
>
```



Stem and Leaf Plot

```
> library(MASS)
> abbey
 [1]   5.2   6.5   6.9   7.0   7.0   7.0
 [7]   7.4   8.0   8.0   8.0   8.0   8.5
[13]   9.0   9.0  10.0  11.0  11.0  12.0
[19]  12.0  13.7  14.0  14.0  14.0  16.0
[25]  17.0  17.0  18.0  24.0  28.0  34.0
[31] 125.0
> round(abbey)   # see help(round) about funny IEEE rounding
 [1]   5   6   7   [ . . . ]
> floor(abbey+.5)  # or round(abbey+.0000001)
 [1]   5   7   7   [ . . . ]
> stem(abbey)

  The decimal point is 1 digit(s) to the right of the |

   0 | 57777778888999011224444 6778
   2 | 484
   4 |
   6 |
   8 |
  10 |
  12 | 5

> stem(abbey[abbey != 125])

  The decimal point is 1 digit(s) to the right of the |

  0 | 57777778888999
  1 | 011224444
  1 | 6778
  2 | 4
  2 | 8
  3 | 4

>
```