R computing language

R is a computing and graphics package for Windows PC, Macintosh, and Linux. It is free and can be downloaded from the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/

R is gaining popularity over other packages such as SAS, Stata, and SPSS because of its price and ease with which software functions can be created and downloaded.

Many specialized statistical techniques have been implemented in R and are available for free download as packages from CRAN.

The goal of statistics

The ultimate goal of statistics is to tell us about the *population* that *data* come from.

The population is infinite, data are not. The population is the ideal; data are an imperfect, hopefully representative snapshot of the population.

Data are random variables

$$\text{data} = X_1, X_2, \ldots, X_n,$$

until they are seen, then they are "observations"

$$X_1 = x_1,\ X_2 = x_2,\ \ldots\ ,\ X_n = x_n.$$

Silly example: say we are interested in the caloric content of poultry hot dog brands. We assume an infinite population of hot dog brands with density $f(x)$. If we know $f(x)$ then we can find, e.g., $x_{0.5} = F^{-1}(0.5)$ and $P(X \geq 100)$ for a randomly drawn poultry hot dog brand.

Our experiment is to go collect $n = 17$ observations. Before we do this, the *data are random*:

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}.$$

After the experiment we have recorded the observed data:

$x_1, \ldots, x_{17} =$

$129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144.$

This is all the information we have to estimate aspects of the unknown density $f(x)$. If this is a "bad" (i.e. unrepresentative or *biased*) sample, all of our estimates will be bad too.

There are two approaches to estimating aspects of $f(x)$: *parametric* and *nonparametric*.

- **Parametric** statistics assumes one member from a particular family of probability models generated $X_1, \ldots, X_{17}$, e.g.

$$X_1, \ldots, X_{17} \overset{iid}{\sim} N(\mu, \sigma^2).$$

Then all questions are answered by first estimating the unknown parameters (μ, σ^2) using the data $X_1, \ldots, X_{17}$. These are denoted $(\hat{\mu}, \hat{\sigma}^2)$. Then the median M is estimated by $\hat{M} = \hat{\mu}$ and

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(x - \hat{\mu})^2\right\}.$$

- **Nonparametric** statistics makes no (rigid) assumption about the shape of $f(x)$, but rather uses the whole of the data to estimate $f(x)$ or $F(x)$. Common nonparametric estimates of $f(x)$ include histograms and kernel smoothers. A common estimate of $F(x)$ is the *empirical distribution function*

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}.$$

The function $\hat{F}(x)$ is simply the proportion of observations less than x, a very natural way of estimating the probability of $P(X \leq x)$. Here $I\{A\} = 1$ if A is true and $I\{A\} = 0$ otherwise. The median is estimated by $\hat{M} = \hat{F}^{-1}(0.5)$, or equivalently, by the *sample median*, the middle value of the ordered observations.

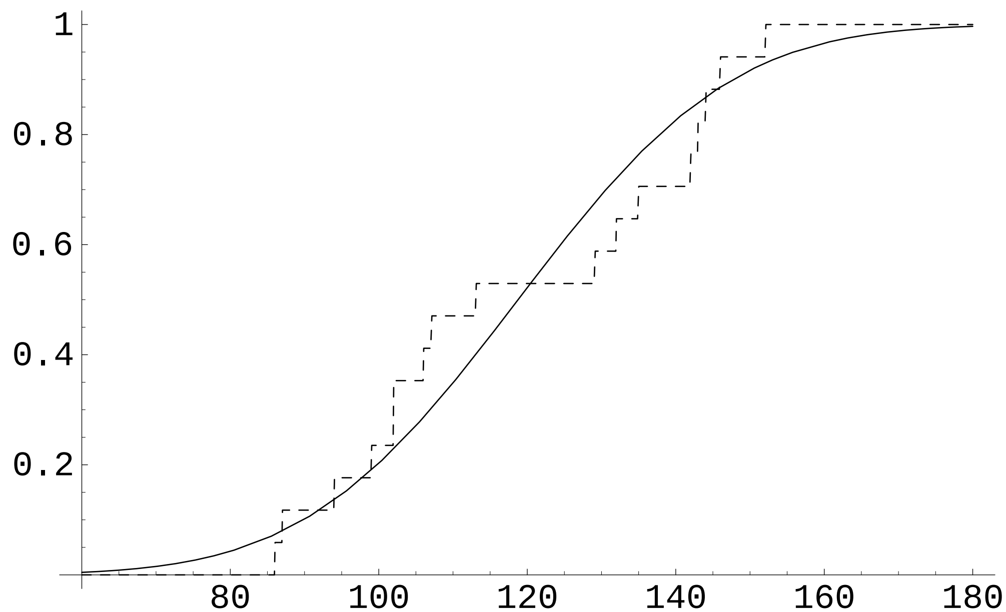

Figure 1: Estimated $F(x)$ for hot dog data. Smooth one assumes $N(118.77, 22.55^2)$, step function is empirical $\hat{F}(x)$.

Assuming normal distribution, $\hat{M} = 118.8$, nonparametric estimate is $\hat{M} = \hat{F}^{-1}(0.5) = 113$.

Looking at data

We say $X_1, \ldots, X_n$ are independent and identically distributed, in shorthand *iid*, if they are independent and all have the same marginal distribution. Then $X_1, \ldots, X_n$ are also called a *random sample*.

We use data to estimate aspects of a *population* that we are interested in.

For example, for continuous outcomes X_i, we may be interested in the population density curve $f(x)$, or quantiles of the population distribution $x_p = F^{-1}(p)$.

Say we've got a random sample $X_1, \ldots, X_n$ of data.

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the values of $\{X_1, \ldots, X_n\}$ ordered from smallest to largest. These are called the *order statistics* of $X_1, \ldots, X_n$.

Ideally, any definition of the p^{th} *sample quantile* $\hat{x}_p$ would require

$$\frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq \hat{x}_p\} \geq p \text{ and } \frac{1}{n}\sum_{i=1}^{n} I\{X_i \geq \hat{x}_p\} \geq (1-p).$$

In words: the proportion of $\{X_1, \ldots, X_n\}$ less than $\hat{x}_p$ is p.

A nice approximation is $\hat{x}_p = X_{(\lceil np \rceil)}$.

Here, $\lceil x \rceil$ is the *ceiling function* of x. It finds the smallest integer that is larger or equal to x. For example $\lceil 2 \rceil = 2$, $\lceil 2.1 \rceil = 3$, and $\lceil 2.9 \rceil = 3$.

Example: The calories in $n = 17$ brands of poultry hot dogs is recorded off the backs of packages. The observed data are $x_1, \ldots, x_{17} =$ $129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144.$ Let's read them into `R` and then find the order statistics.

```
> cal <- c(129,132,102,106,94,102,87,99,107,113,135,142,86,143,152,146,144)
> cal.ordered <- sort(cal)
> cal.ordered
 [1]  86  87  94  99 102 102 106 107 113 129 132 135 142 143 144 146 152
```

Using the definition $\hat{x}_p = X_{(\lceil np \rceil)}$, the sample median is given by

```
> cal.ordered[ceiling(17*0.5)]
[1] 113
```

Let's see what R's built-in `quantile()` function gives us:

```
> quantile(cal,0.5)
50%
113
```

This estimates the *unknown* population median $x_{0.5} = F^{-1}(0.5)$.

Now let's try the 90^{th} percentile:

```
> cal.ordered[ceiling(17*0.9)]
[1] 146
> quantile(cal,0.9)
  90%
144.8
```

R does something a bit fancier than our simpler approach (it interpolates). In fact, R has 9 different ways of computing an estimate of x_p! Type `> help(quantile)` to see a description of each.

In practice, with a reasonable amount of data, the different estimates will give almost the same number. If you have a very small amount of data, it's hard to estimate an unknown population quantile x_p unless you assume something stronger, e.g. that the data are $N(\mu, \sigma^2)$, $\exp(\lambda)$, etc.

It is important to stress the difference between x_p and $\hat{x}_p$.

- x_p is *unknown* population quantile. We'd like to know what it is but have to guess.
- $\hat{x}_p$ is a guess at x_p based on what we *can* observe: some data $X_1, X_2, \ldots, X_n$. In fact as n gets large, $\hat{x}_p \to x_p$. We'll formalize what this means later.

Boxplots

Textbook: Section 10.6 (read pp. 401–404).

The quantiles can be used graphically to describe how data are spread out using a boxplot.

Let $\hat{x}_{0.25}$, $\hat{x}_{0.5}$, and $\hat{x}_{0.75}$ be the first, second, and third quartiles of the data (the second quartile is also the sample median). These estimate their population analogues and form the "box" in the boxplot.

The $\mathrm{IQR} = \hat{x}_{0.75} - \hat{x}_{0.25}$ is the length of the interval encompassing 50% of the sample. The inner fences are located at $\hat{x}_{0.25} - 1.5\ \mathrm{IQR}$ and $\hat{x}_{0.75} + 1.5\ \mathrm{IQR}$. Sample points lying outside the inner fences are considered to be outlying. The "whiskers" are lines that extend to the two observations furthest away from $\hat{x}_{0.5}$ within the inner fences.

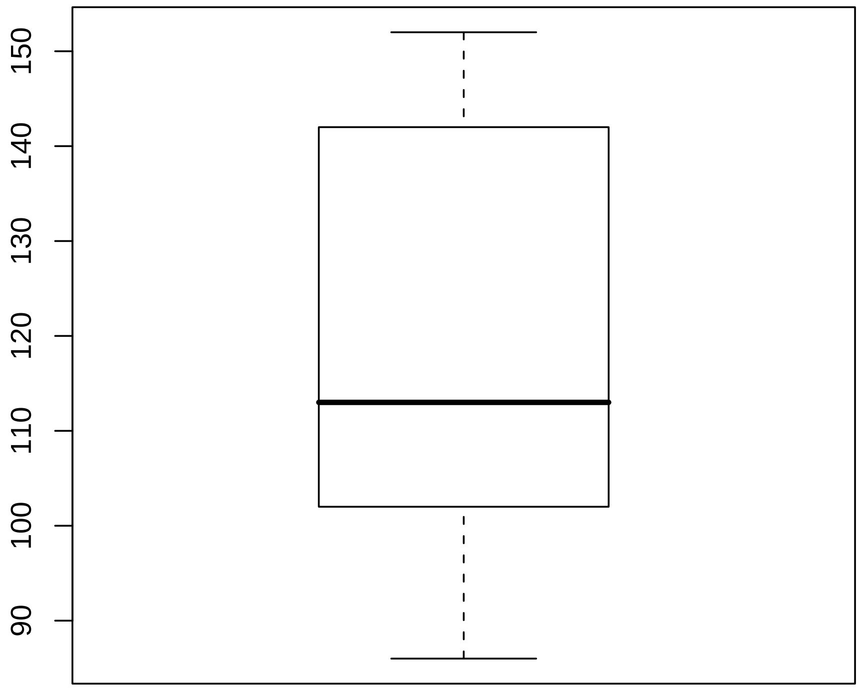

Figure 2: Obtained from typing `> boxplot(cal)` in `R`.

Boxplots, *histograms*, and *normal probability plots* are three ways to visually assess whether a random sample $X_1, \ldots, X_n$ are reasonably (or approximately) normal.

`R` shows outlying points as circles. For *perfectly* normal data, the probability of seeing an outlier is 0.0070. For a sample size 150, in perfectly normal data we *expect* to see $0.007 \times 150 \approx 1$ outlier. In a sample size of 1000 we expect to see roughly 7 outliers and so on. If you see *no* outliers in a sample of 1000, then the data probably aren't normal, even if the histogram looks symmetric and bell shaped. If you see 10 outliers in a sample of size 100 your data probably aren't normal.

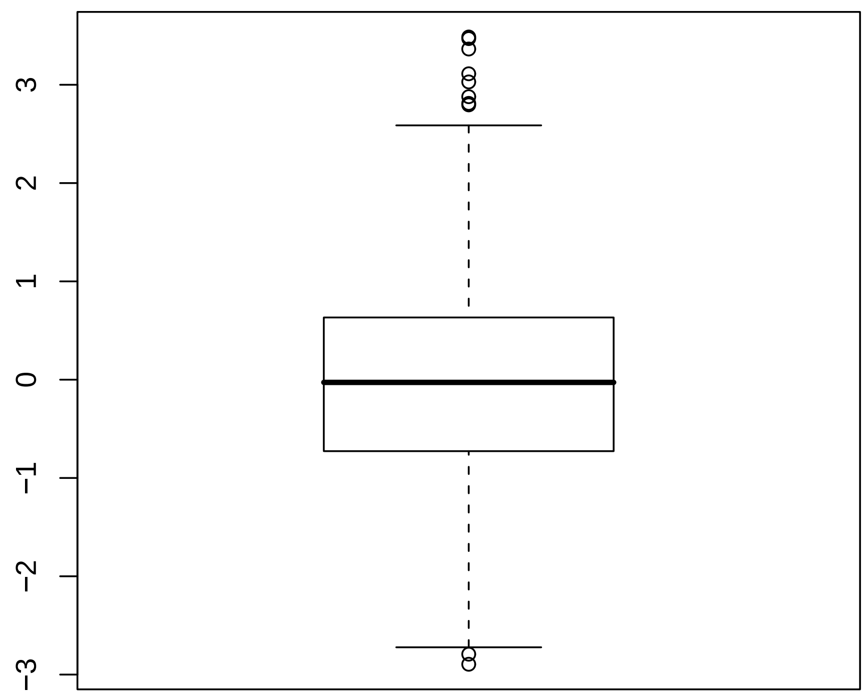

Figure 3: `> d <- rnorm(1000); boxplot(d)`. $X_1, \ldots, X_{1000} \overset{iid}{\sim} N(0,1)$.

• The boxplot should reflect symmetry, if present, in the population density in terms of the box and the whiskers.

• When checking for normality, histograms will never look perfectly normal. When sample sizes are small, say 20 or less, the boxplot is more useful than the histogram.

• When checking for normality, we expect a boxplot to be *roughly* symmetric and to not have a very large number of outliers.

• The boxplot cannot tell you anything about multimodality; recall that the median (and mean) can occur in areas of very low probability mass. So can the quartiles. Histograms do a better job of showing how probability mass is actually spread out.

Histograms

Textbook: Figures 3.3 (p. 74) & 10.8 (p. 390), Section 10.3.

Data might be *approximately* distributed $N(\mu, \sigma^2)$, Weibull(α, λ), etc., but never exactly.

A histogram provides a "nonparametric" estimate of an unknown density $f(x)$ based on a random sample $X_1, \ldots, X_n$ and some intervals, termed "bins."

The histogram $\hat{f}(x)$ is a step function that integrates to one. Let $a_0 < a_1 < a_2 < \cdots < a_{m-1} < a_m$ be interval endpoints defining m bins that data fall into. We must have $a_0 < X_i < a_m$ for all $i = 1, \ldots, n$. Then

$$\hat{f}(x) = \frac{1}{(a_j - a_{j-1})n} \sum_{i=1}^{n} I\{a_{j-1} \leq X_i < a_j\} \text{ for } a_{j-1} \leq x < a_j.$$

Let's verify that the area under $\hat{f}(x)$ is one. Let n_j be the number of $\{X_1, \ldots, X_n\}$ between a_{j-1} and a_j. So $n_1 + n_2 + \cdots + n_m = n$.

$$\begin{aligned}
\int_{a_0}^{a_m} \hat{f}(x)dx &= \sum_{j=1}^{m} \left(\frac{n_j}{n}\right) \left(\frac{a_j - a_{j-1}}{a_j - a_{j-1}}\right) \\
&= \frac{1}{n} \sum_{j=1}^{m} n_j = \frac{n}{n} = 1.
\end{aligned}$$

```
> par(mfrow=c(2,2))
> x <- seq(-3.5,3.5,0.1); y <- dnorm(x,0,1); data <- rnorm(200,0,1)
> hist(data,freq=FALSE,xlim=c(-3.5,3.5),ylim=c(0,0.6)); lines(x,y)
> x <- seq(-3.5,3.5,0.1); y <- dnorm(x,0,1); data <- rnorm(200,0,1)
> hist(data,freq=FALSE,xlim=c(-3.5,3.5),ylim=c(0,0.6)); lines(x,y)
> x <- seq(-3.5,3.5,0.1); y <- dnorm(x,0,1); data <- rnorm(200,0,1)
> hist(data,freq=FALSE,xlim=c(-3.5,3.5),ylim=c(0,0.6)); lines(x,y)
> x <- seq(-3.5,3.5,0.1); y <- dnorm(x,0,1); data <- rnorm(200,0,1)
> hist(data,freq=FALSE,xlim=c(-3.5,3.5),ylim=c(0,0.6)); lines(x,y)
```

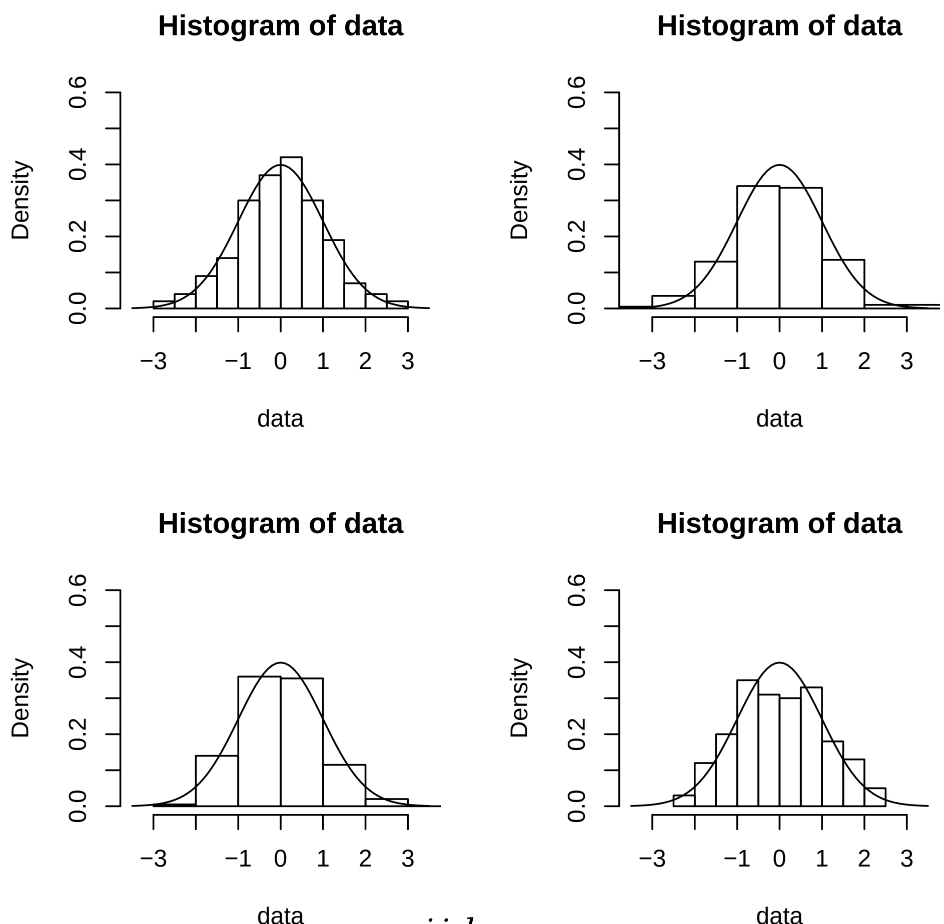

Figure 4: $X_1, \ldots, X_{200} \overset{iid}{\sim} N(0,1)$ repeated four times.

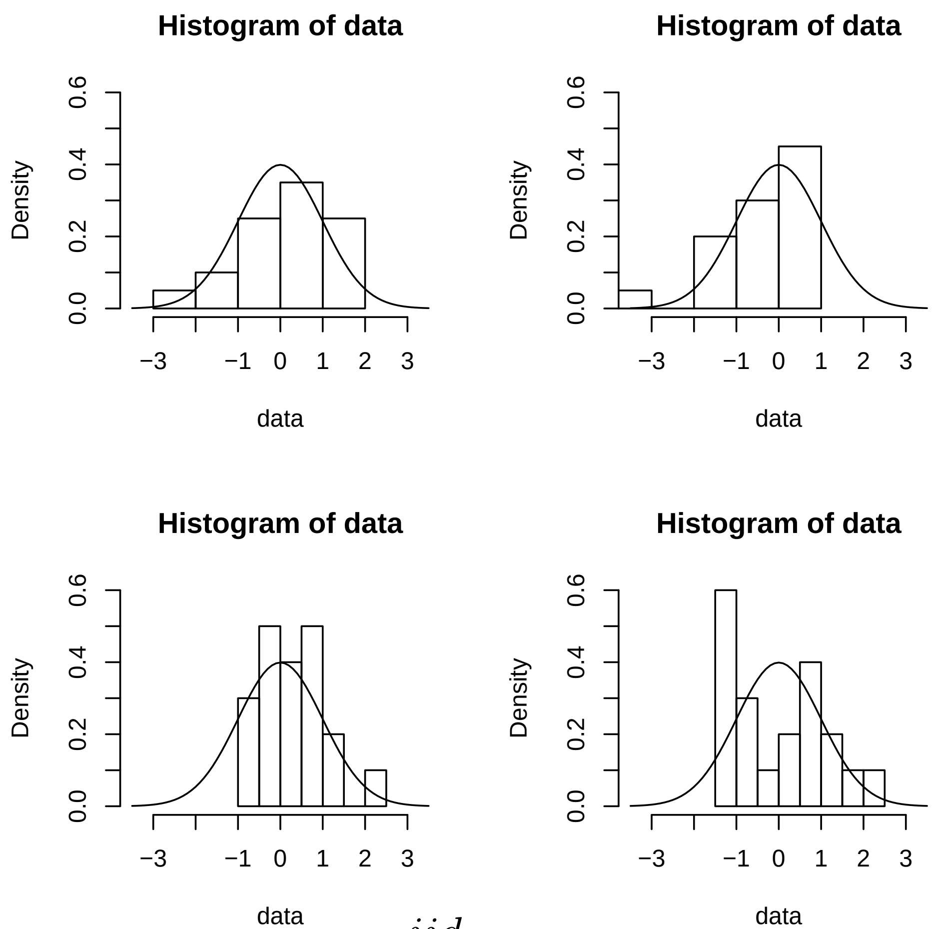

Figure 5: $X_1, \ldots, X_{20} \overset{iid}{\sim} N(0,1)$ repeated four times.

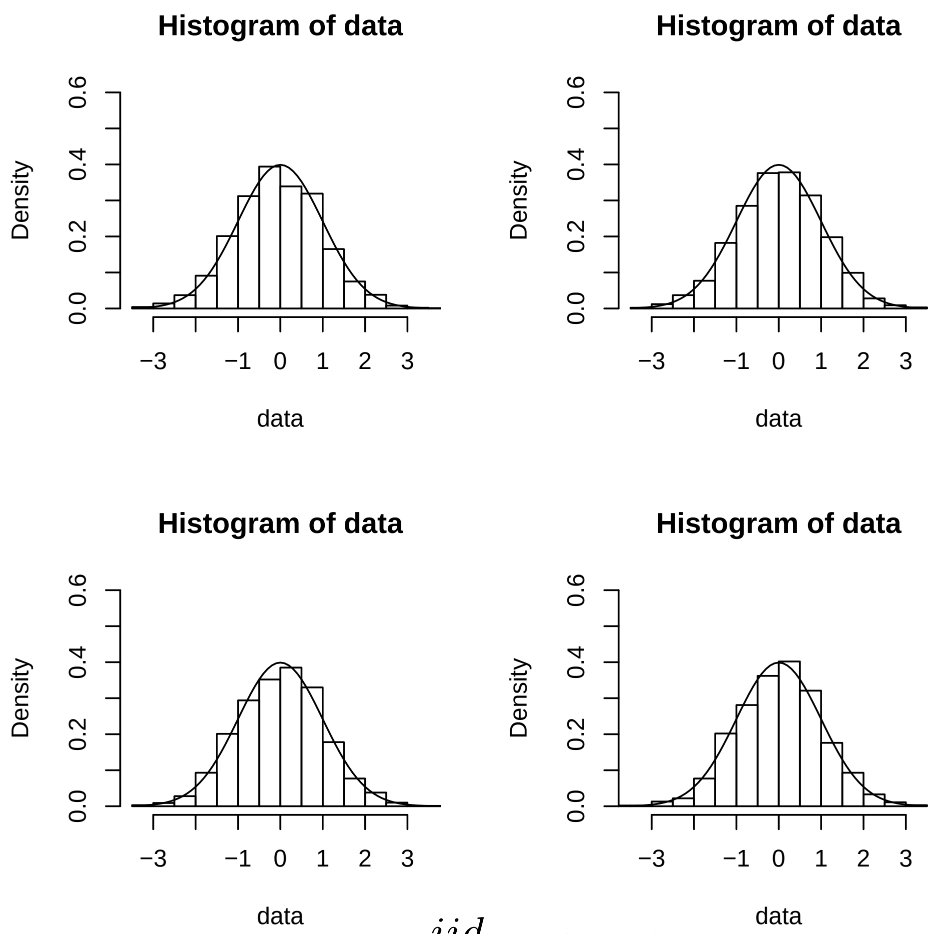

Figure 6: $X_1, \ldots, X_{2000} \overset{iid}{\sim} N(0,1)$ repeated four times.

- As n gets large, the histogram $\hat{f}(x)$ better estimates the true $f(x)$.
- These data are *perfectly* normal, but because our information is limited, e.g. $n = 20$, $n = 200$, or $n = 2000$ observations, our estimate $\hat{f}(x)$ is also imperfect.
- However, we usually only get *one* data set $X_1, \ldots, X_n$ to look at! If a histogram is "non-normal" looking, the data still may be approximately normal. You are looking for *extreme* skewness, tons of outliers, or marked multimodality to call data "non-normal."
- An aside: $(n_1, n_2, \ldots, n_m)$ are jointly distributed and have a multinomial distribution before data $X_1, \ldots, X_n$ are collected and actually seen (pp. 73–75).

Recap:

- $X_1, \ldots, X_n$ are a random sample from an *unknown* pdf $f(x)$.
- $\hat{x}_p$ estimates the unknown p^{th} quantile x_p, defined as $P(X_i \leq x_p) = p$. Number such that the sample proportion of data $\{X_1, \ldots, X_n\}$ less than $\hat{x}_p$ is about p.
- A boxplot is a graphical simplification of the data using the sample quartiles $\hat{x}_{0.25}$, $\hat{x}_{0.5}$, and $\hat{x}_{0.75}$, as well as "inner fences" that contain "most" of the probability mass.
- A histogram is a simple estimate of $f(x)$ constructed by estimating the probability of data falling into bins.

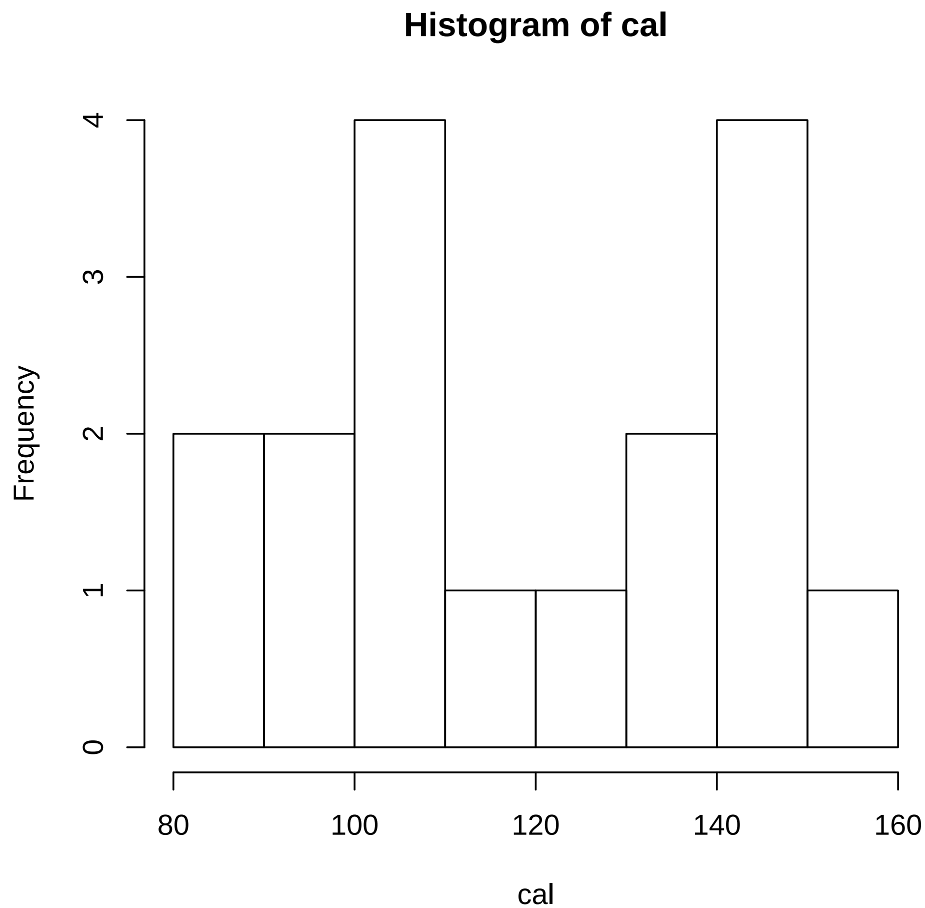

Figure 7: Calories in different brands of poultry hot dogs. $(x_1, \ldots, x_{17})$ is equal to $(129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144)$. This histogram is the raw frequency counts, the default in `R`.

4.3 Covariance and correlation

The variance of X is a measure of how "variable" or spread out outcomes X from the experiment are. The *covariance* between X and Y is a crude measure of related outcomes X and Y are.

def'n: The covariance σ_{XY} between any X and Y is defined to be

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E\{(X - E(X))(Y - E(Y))\},$$

where this expectation is with respect to the joint distribution of (X, Y).

Let μ_X and μ_Y be the marginal means of X and Y and let σ_X^2 and σ_Y^2 be the marginal variances.

Prop: $\mathrm{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$.

We'll prove this on the board.

For (X, Y) with pdf $f(x, y)$ we have

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dxdy - \mu_X \mu_Y,$$

where

$$\mu_x = \left[\int_{-\infty}^{\infty} xf_X(x)dx \right] \text{ and } \mu_Y = \left[\int_{-\infty}^{\infty} yf_Y(y)dx \right].$$

Interpretation: if X and Y tend to increase/decrease together as a pair then $\text{Cov}(X, Y) > 0$, if X tends to increase when Y decreases or vice-versa then $\text{Cov}(X, Y) < 0$.

The units for the covariance are the units of X times the units for Y, not very interpretable. A standardized version of the covariance is the correlation.

def'n: The correlation between any X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

The correlation is a unitless measure of linear association between X and Y. Note that $\rho_{XY} = 0$ iff $\mathrm{Cov}(X,Y) = 0$ and we say X and Y are uncorrelated. This *does not* imply that X are Y are not related, i.e. independent.

Prop: $-1 \leq \rho \leq 1$. Proof in text.

Interpretation: When $\rho_{XY} > 0$, X and Y are positively linearly associated. In fact, when $\rho_{XY} = 1$, $X = a + bY$ for some fixed a and $b > 0$. That is, $\rho_{XY} = 1$ implies that X and Y are *perfectly* linearly associated. Similarly, $\rho_{XY} < 0$ implies that X and Y are negatively linearly associated and $\rho_{XY} = -1$ forces $X = a + bY$ where $b < 0$.

Useful result: the marginal mean $E(X)$ can be obtained from jointly distributed (X, Y) as, e.g. in the continuous case,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dydx \\ &= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)dy\right]dx \\ &= \int_{-\infty}^{\infty} xf_X(x)dx. \end{aligned}$$

We are treating this expectation $E(X)$ as initially being with respect to (X, Y) and $g(x, y) = x$. Recall that in general

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dydx.$$

A similar result holds for $E(Y)$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, discrete random pairs, & can be generalized to random vectors $\mathbf{X} = (X_1, \ldots, X_n)$.

Example: Let (X, Y) be jointly distributed with pdf

$$f(x,y) = \left\{ \begin{array}{ccc} x - y & \text{for} & \begin{array}{c} 1 \le x \le 2 \\ 0 \le y \le 1 \end{array} \\ & & \\ 0 & & \text{otherwise} \end{array} \right\}.$$

Are X and Y independent? Why or why not? Find the conditional pdf of X given that $Y = \frac{1}{2}$, $f_{X|Y}(x|\frac{1}{2})$. Find ρ_{XY}.

No, $f(x, y)$ cannot be factored into $f(x, y) = h(x)g(y)$ so X and Y are not independent.

To get $f_{X|Y}(x|y)$ we first need $f_Y(y)$:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_1^2 (x-y)dx = 3/2 - y,$$

on $R_Y = (0,1)$. So

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(t)} = \frac{x-y}{3/2-y},$$

and $f_{X|Y}(x|0.5) = x - 0.5$ on $R_{X|Y=0.5} = (0,1)$.

Recall $\sigma_{XY} = E(XY) - \mu_X\mu_Y$.

$$\mu_X = E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dxdy = \int_0^1\int_1^2 x(x-y)dxdy = 19/12.$$

$$\mu_Y = E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dxdy = \int_0^1\int_1^2 y(x-y)dxdy = 5/12.$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \int_0^1\int_1^2 xy(x-y)dxdy = 2/3.$$

So

$$\sigma_{XY} = E(XY) - E(X)E(Y) = 2/3 - (19/12)(5/12) = 1/144.$$

$$\begin{aligned}\sigma_X^2 = E((X-E(X)^2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-19/12)^2 f(x,y)dxdy \\ &= \int_0^1\int_1^2 (x-19/12)^2(x-y)dxdy = 11/144.\end{aligned}$$

$$\begin{aligned}\sigma_Y^2 = E((Y-E(Y)^2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-5/12)^2 f(x,y)dxdy \\ &= \int_0^1\int_1^2 (y-5/12)^2(x-y)dxdy = 11/144.\end{aligned}$$

Finally,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{1/144}{\sqrt{(11/144)(11/144)}} = \frac{1}{11} \approx 0.09.$$

We see a moderate, positive relationship between X and Y.

Prop: If X is independent of Y then $\rho_{XY} = \sigma_{XY} = 0$.

Proof: $\sigma_{XY} = E(XY) - \mu_X \mu_Y = E(X)E(Y) - \mu_X \mu_Y = 0$.

Prop: For any (X, Y),

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

We'll prove this on the board.

Prop: If X is independent of Y then
$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. (Immediate from previous props).

Theorem A, Corollary A, and Corrolary B (p. 140) are generalizations of these simpler propositions.

Note: By definition, $\text{Cov}(X, X) = E((X - E(X))^2) = \text{Var}(X) = \sigma_X^2$.

For previous example, what is $\text{Var}(X + Y)$?

ρ_{XY} is typically not known but rather estimated from data $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ directly by

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

This estimate does not come from any probability model, but does correspond to the MLE for bivariate normal data (more to come).

For the previous example, a lot of calculus was required to get the answer. Another approach is to simulate a lots of (X_i, Y_i) from the joint pdf $f(x, y)$ and compute the sample correlation coefficient. We'll discuss later how $\hat{\rho} \xrightarrow{P} \rho$ as n gets big.

Mathematica code to implement a Metropolis-Hastings sampler:

```
n=20000; t=Table[{0,0},{n}]; t[[1]]={1.5,0.5};
For[i=2,i<=n,i++,
    x=Random[]+1.0; y=Random[];
    If[Random[]<(x-y)/(t[[i-1,1]]-t[[i-1,2]]),
      t[[i]]={x,y},t[[i]]=t[[i-1]]]];
CorrelationMatrix[t][[1,2]]
0.091422
```

$\hat{\rho} = 0.0914$ is close to $\rho = 0.0909$.

Example: Here's the Gesell score y_i and age at first word in months x_i data, $i = 1, \ldots, 21$.

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
15	95	26	71	10	83	9	91	15	102
20	87	18	93	11	100	8	104	20	94
7	113	9	96	10	83	11	84	11	102
10	100	12	105	42	57	17	121	11	86
10	100								

We compute $\hat{\rho}_{XY} = -0.640$, a moderately strong negative relationship between age at first word spoken and Gesell score. Note that the correlation is the same when age is measured in years instead of months.

```
> age<-c(15,26,10,9,15,20,18,11,8,20,7,9,10,11,11,10,12,42,17,11,10)
> Gesell<-c(95,71,83,91,102,87,93,100,104,94,113,96,83,84,102,100,105,57,121,86,100)
> plot(age,Gesell)
> cor(age,Gesell)
[1] -0.64029
```

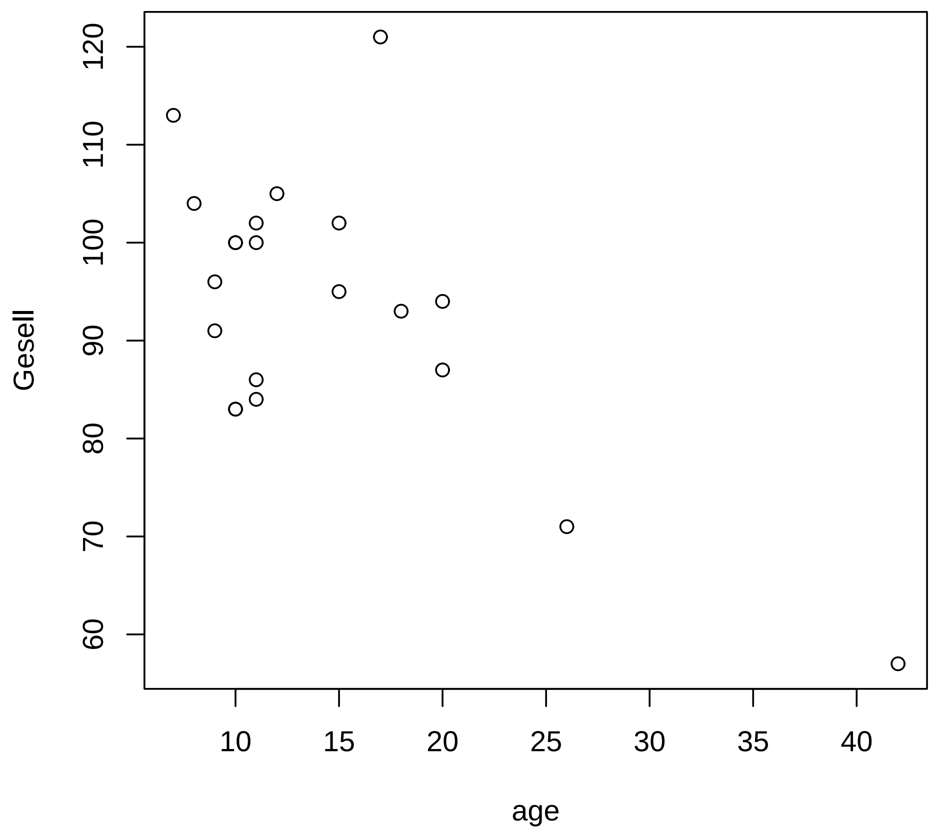

Figure 8: Scatterplot of $(x_1, y_1), \ldots, (x_{21}, y_{21})$.