Statistical Models, Sampling and Sampling Distributions

MAST90044

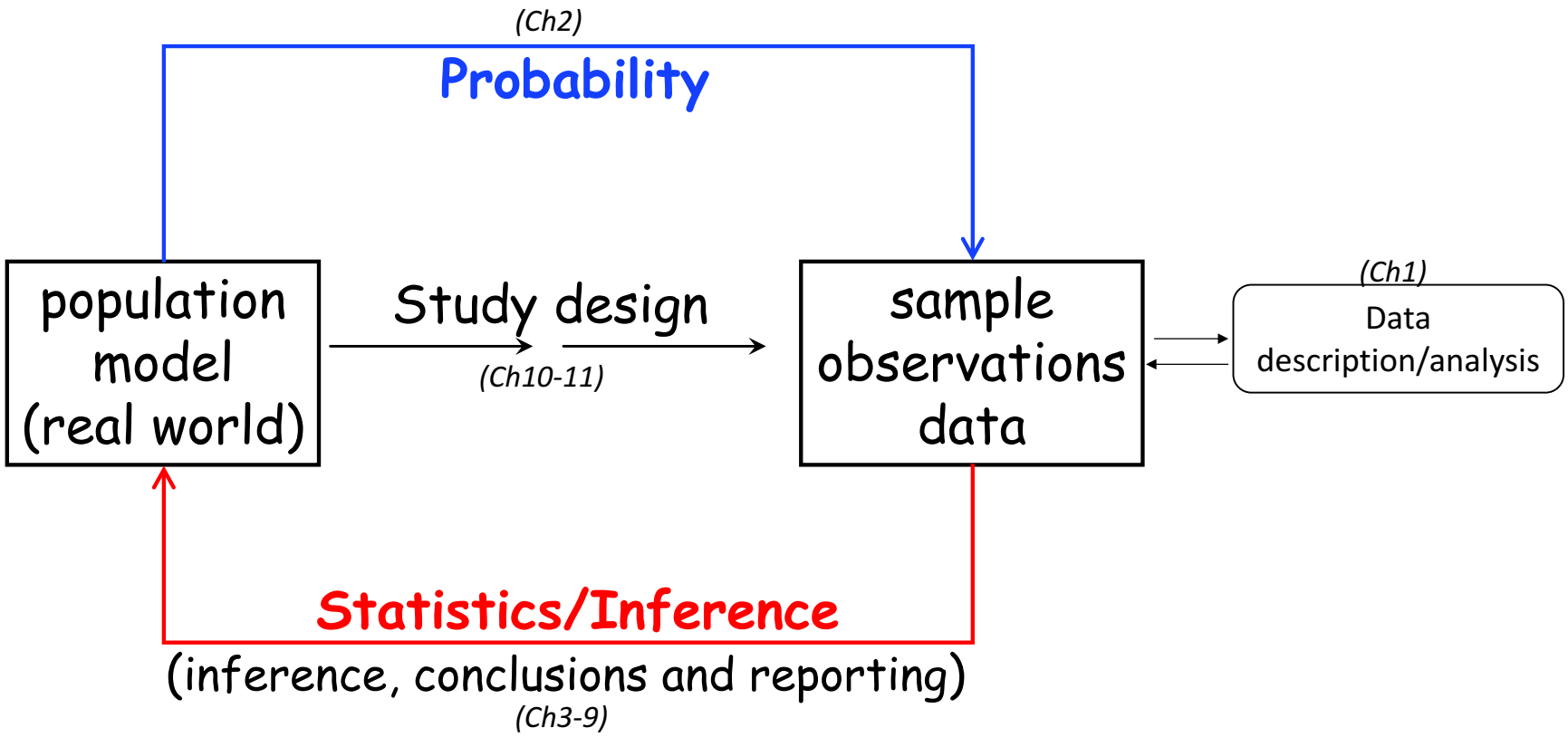
Thinking and Reasoning with Data

Dr Julia Polak

Chapter 2 – Lecture 4 & 5

Department of Mathematics & Statistics
University of Melbourne

Probability and Statistics



Outline

Statistical Models (model variation in the data, used to make inferences)

Sampling (how to select data)

Sampling Distributions (model sample-to-sample variation)

Statisitcal Models

We all use models.

Models have assumptions.

Assumptions should be checked.

All models are wrong, but some are useful.

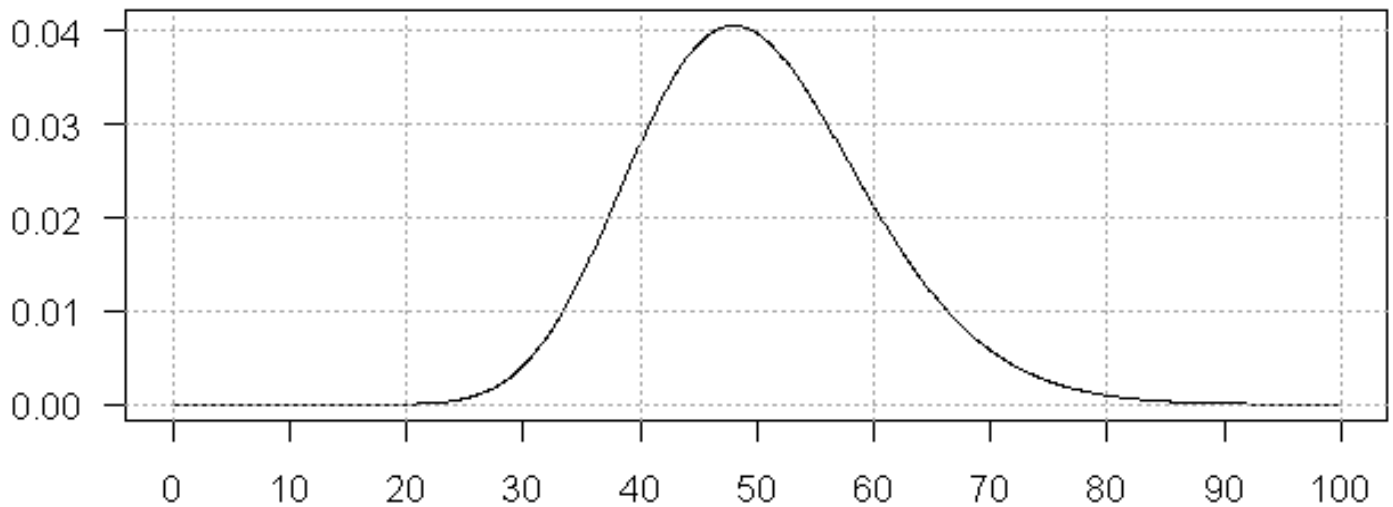
George Box.

Sampling Model

The most fundamental statistical model is that of "random sampling from an infinite population".

(What is the difference between $\bar{X}$, $\bar{x}$ and μ?)

population (model) ($\overset{\circ}{\underset{\circ}{\circ}}$)



distribution $\mathbb{D}$; mean μ (50); sd σ (10)
(parameters)

Sampling Model - Example

observations ($n = 20$)

$$X_1, \quad X_2, \quad \ldots, \quad X_{20}. \qquad \text{“BEFORE”}$$

(iidrvs) (population)

$$x_1 = 47.3, \quad x_2 = 61.2, \quad \ldots, \quad x_{20} = 41.6. \qquad \text{“after” (sample)}$$

sample mean $\bar{x} = \frac{1}{20}(47.3 + 61.2 + \cdots + 41.6) = 50.87$ “after”

$\bar{x}$ a realisation of $\bar{X} = \frac{1}{20}(X_1 + X_2 + \cdots + X_{20})$. “BEFORE”

$\bar{X}$ a random variable: it has a (theoretical) distribution.
The empirical distribution is constructed from the observed data i.e. the sample.

Random Variables

A random variable is a numerical outcome of a "random" process. Here "random" means uncertain. Before the procedure is conducted and we make the observation, we do not know what its value will be e.g. $X, \bar{X}$. After the procedure we obtain the *observations* e.g. $x, \bar{x}$.

For example:

1. treating ten individuals: X = number cured (discrete)
2. patient diagnosed with cancer: Z = survival time (continuous)
3. measuring blood pressure: U = pressure reading (continuous)

The set of possible values of the random variable ($X, Z, U, \bar{X}$, etc.) is called the sample space.

Random Variables

We consider two types of random variables:

- Discrete variables can only take certain values. Usually they are based on counts.
  e.g. number of children in a family.
- Continuous variables can take any value in an interval (or set of intervals). Usually these are measurements of some kind.
  e.g. blood pressure.

Probability distributions

For any random variable (e.g. $X, \bar{X}$), there is a rule (or set of rules) which governs which values the variable can take, and how likely each of the various values are.

For example, if we are sampling from a population of people, the random variable "age" will depend on the demographics of the population.

We call this the distribution of the random variable.

The distribution of a random variable is defined by a function.

Random Variables and Probability Distributions

In a long run of repeated samples, the value of the random variable (RV) is thought to follow some rule of probability, which may be described by some mathematical relationship. This defines the *(theoretical) distribution of the random variable*.

If you took many samples would you expect the same result?

Probability Models

...describe the distribution of a population.

We use probability models to model the variation in outcomes of a random process and decide whether or not the observed data is compatible with the model... or whether the model is compatible with the data.

Is what we observed (in the data) what we would expect to observe (according to the probability model) if the model was correct?

What type of probability model would help us to model the outcome of a random process, e.g.:

- number of individuals cured by a treatment?
- number of weeds found in a plot of land?
- number of contaminated bananas in a shipment?

Sampling Model - Example

observations ($n = 20$)

$$X_1, \quad X_2, \quad \ldots, \quad X_{20}. \qquad \text{“BEFORE”}$$

(iidrvs) (population)

$$x_1 = 47.3, \quad x_2 = 61.2, \quad \ldots, \quad x_{20} = 41.6. \qquad \text{“after” (sample)}$$

sample mean $\bar{x} = \frac{1}{20}(47.3 + 61.2 + \cdots + 41.6) = 50.87$ “after”

$\bar{x}$ a realisation of $\bar{X} = \frac{1}{20}(X_1 + X_2 + \cdots + X_{20})$. “BEFORE”

$\bar{X}$ a random variable: it has a (theoretical) distribution.
The empirical distribution is constructed from the observed data i.e. the sample.

Sampling Model - Example

$$\mathrm{E}(\bar{X}) = \mu \qquad (50) \qquad \textit{(distn of } \bar{X} \textit{ will be centered about } \mu\textit{)}$$

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n} \qquad (\frac{10^2}{20})$$

Thus $\bar{x}$ is an unbiased estimate of μ *(the mean of the sampling distn is equal to the parameter)*;

and its accuracy (precision) is measured by $\mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

and this is true no matter what the population distribution!

But wait, there's more . . .

$$\bar{X} \overset{\mathrm{d}}{\approx} \mathrm{N}(\mu, \frac{\sigma^2}{n}) \qquad \text{for large } n$$

Expectation

The expectation, or mean, of a random variable X is a measure of the "centre" of its probability distribution.

It is a weighted average of the values that X can take, where the weights are provided by the distribution of X.

It is the "centre of mass" of the pmf (discrete vars) or pdf (continuous vars).

It is defined by:

$$\mathrm{E}(X) = \sum xp(x) \quad \text{or} \quad \mathrm{E}(X) = \int xf(x),$$

discrete or continuous.

Notes for the Mean

- E(X) is also written μ or μ_X.
- If the pmf (discrete vars) or pdf (cont. vars) is symmetrical, the mean is on the axis of symmetry.
- The mean is not the "most expected" value (what does that mean?).

  It is not the value we expect to observe; nor the most likely value or even need be near it but is often "around about" its mean.
- The expectation is additive: $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ always!

  "The expectation of a sum is the sum of the expectations"
- $\mathrm{E}(a + bX) = a + b\mathrm{E}(X)$.

Variance and standard deviation

The mean is a good measure of the location of a distribution. But we also want to know other things.

The next thing we need to know is the spread of the distribution. The best measure of this is variance (and standard deviation), which is defined by:

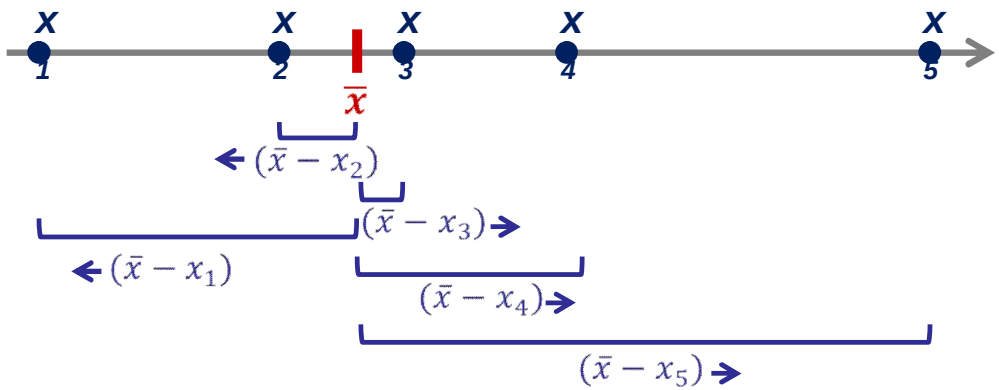
$\sigma^2 = \mathrm{var}(\mathrm{X}) = \mathrm{E}[(\mathrm{X} - \mu)^2]$, where $\mu = \mathrm{E}(\mathrm{X})$

$\sigma = \mathrm{sd}(\mathrm{X}) = \sqrt{\mathrm{var}(\mathrm{X})}$

The more widespread the values of X, the larger $(\mathrm{X} - \mu)^2$ is, and so the larger $\mathrm{var}(\mathrm{X})$.

Variance and standard deviation

Variance



Variance =

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

- Determine the spread of distribution
- Difference of individual observations from mean

  (deviations) = $(x_i - \bar{x})$
- Sum of deviations, squared = $\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Number of independent deviations = (n-1)

Statistical Models – compartmentalise variation in the data

$$\begin{aligned}
\text{response} &= \text{systematic component} &&+ \text{random error} \\
\text{data} &= \text{deterministic component} &&+ \text{stochastic component} \\
&= \text{signal} &&+ \text{noise} \\
&= \text{(explained)} &&+ \text{(unexplained)}
\end{aligned}$$

Our task is to separate the two components.

$$\begin{array}{rcccc}
y & = & \mu(x) & + & e \\
\uparrow & & \uparrow & & \uparrow \\
\text{response} & & \text{explanatory} & + & \text{random} \\
\text{variable} & & \text{variables} & & \text{error}
\end{array}$$

Statistical Models – compartmentalise variation in the data

$$
\begin{array}{ccccc}
y & = & \mu(x) & + & e \\
\uparrow & & \uparrow & & \uparrow \\
\text{response} & & \text{explanatory} & + & \text{random} \\
\text{variable} & & \text{variables} & & \text{error}
\end{array}
$$

	μ
categorical	e.g. μ_f, μ_m
numerical	e.g. $\beta_0 + \beta_1 x$

PARAMETERS: μ; μ_f, μ_m; β_0, β_1; ... & σ

Example: Simon Newcomb and the speed of light

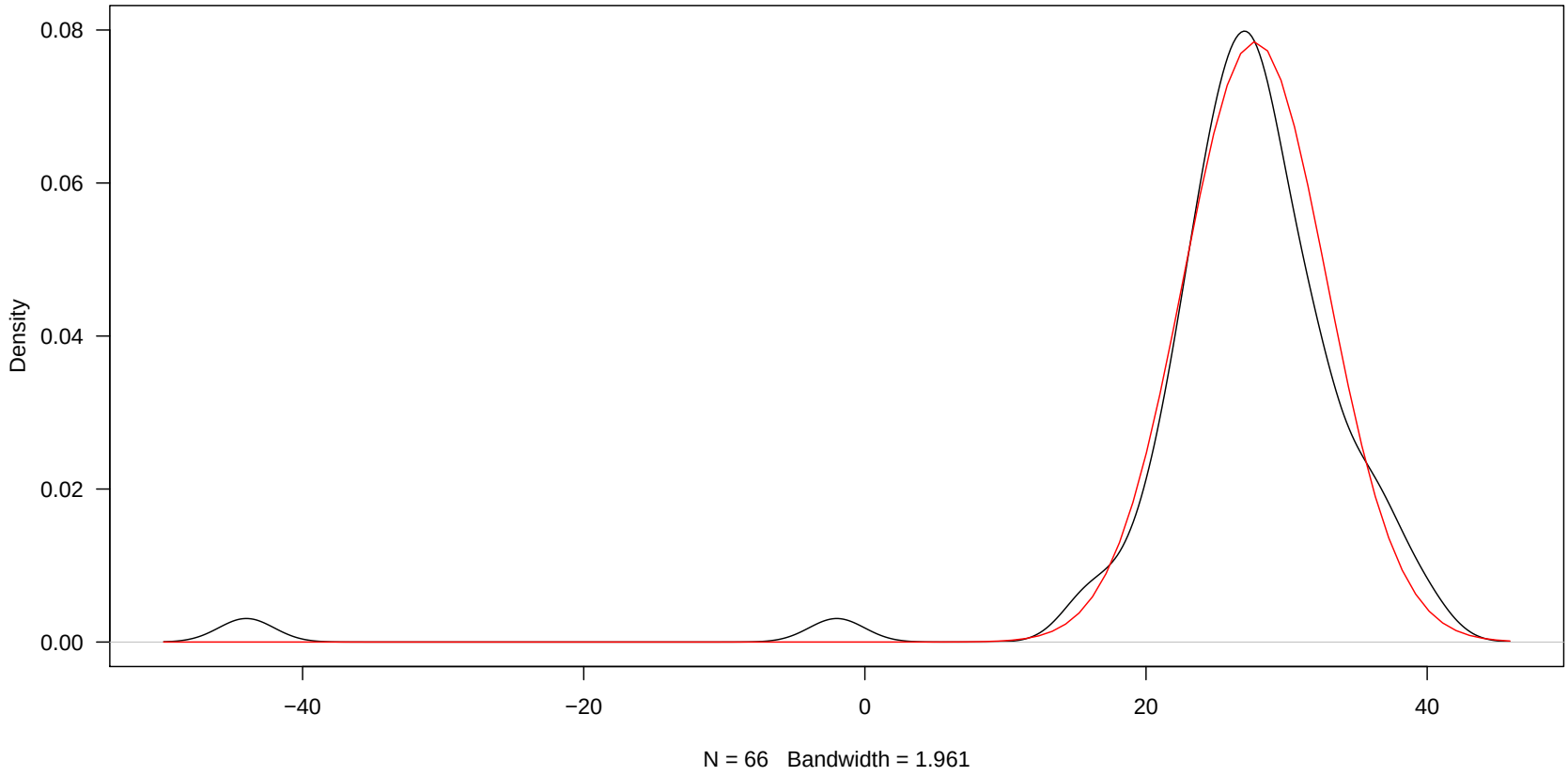
- A series of 66 measurements between July and September 1882.
- Time for a light signal took to pass from his laboratory on the Potomac River to a mirror at the Washington Monument and back (total distance = 7400m).
- First measurement was 0.000024828 seconds = 24,828 nanoseconds.
- Data are deviations from 24,800 nanoseconds.

(Recall: speed = distance/time)

Newcomb's measures of the speed of light

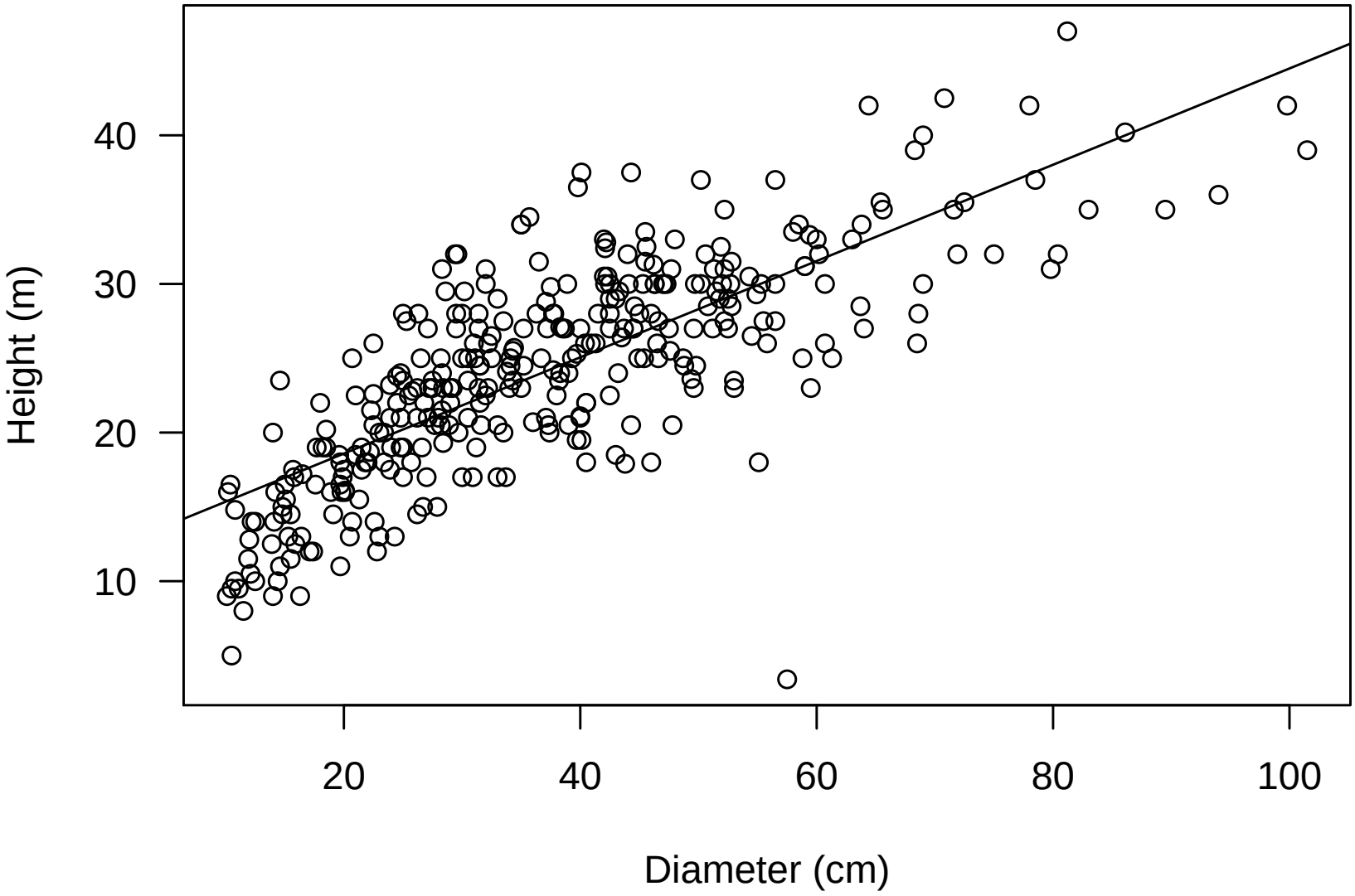
28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

Newcomb's measures of the speed of light



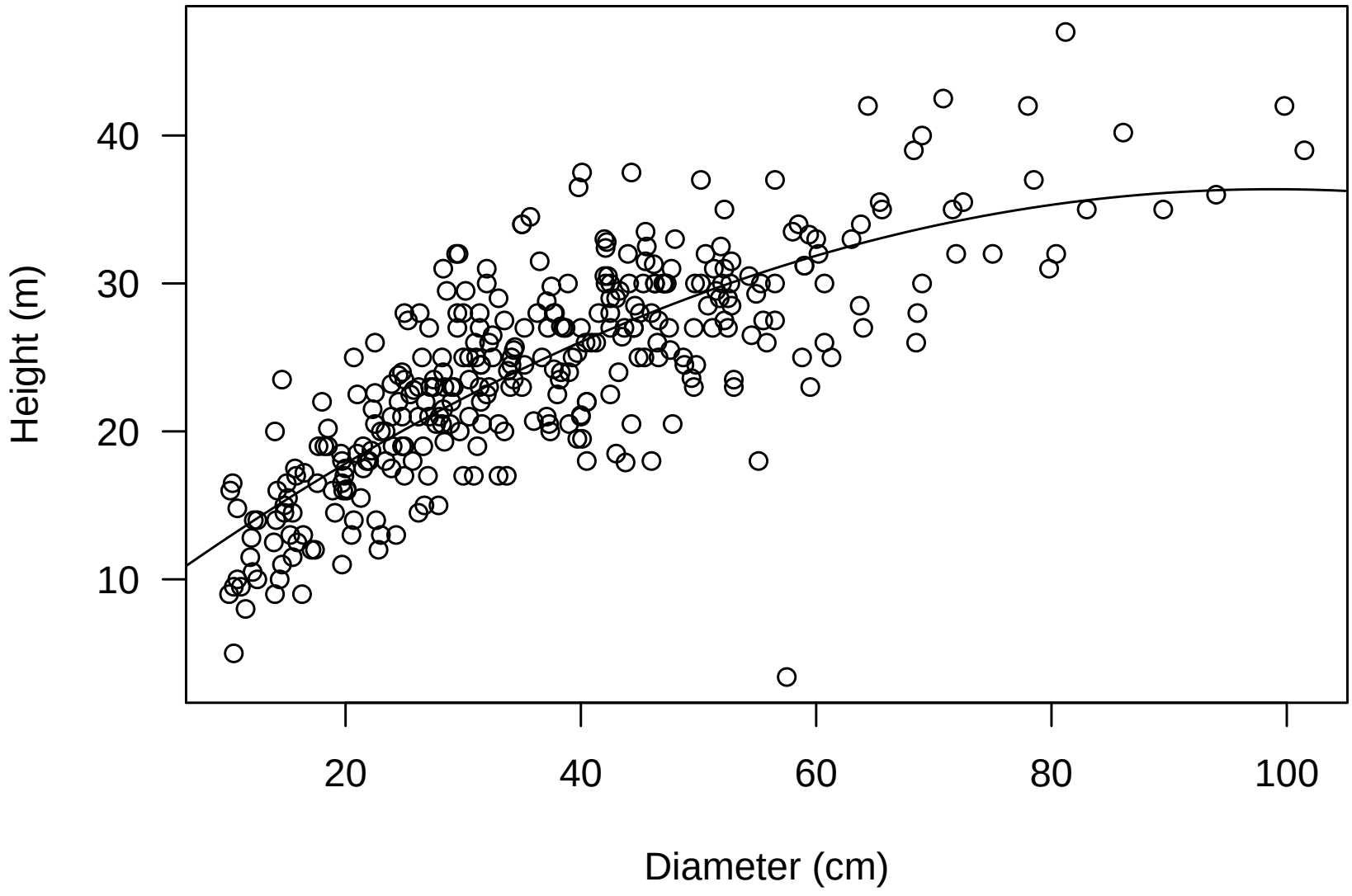
black line: original data with smoother
red line: -44 and -2 removed.
speed of light: 299 792 km/sec.

height vs diameter of trees: straight line model



is this a good fit? is there a better fit?

height vs diameter of trees: quadratic model



try non-linear models e.g. exponentials? not simply cubics, polynomials etc.

Statistical models

Parameters and estimates

- **Parameters** are the true values for the population, and are unknown.
- e.g. true speed of light.
  p; μ; σ; α.
- **Parameter estimates** come from the data.
- e.g. Newcombe's experimental results give *an* estimate of the "true" speed of light.
  $\hat{p} = 0.32$; $\hat{\mu} = 5.7$.

Statistical models – most basic model: Example

No explanatory variables gives the most basic model

Cotton lint yields (in kg per hectare) for 25 farms in the Blackland Prairie:

289 267 305 340 400 380 285 417 351 358 259 294
402 322 348 266 293 415 319 355 389 305 316 331 327

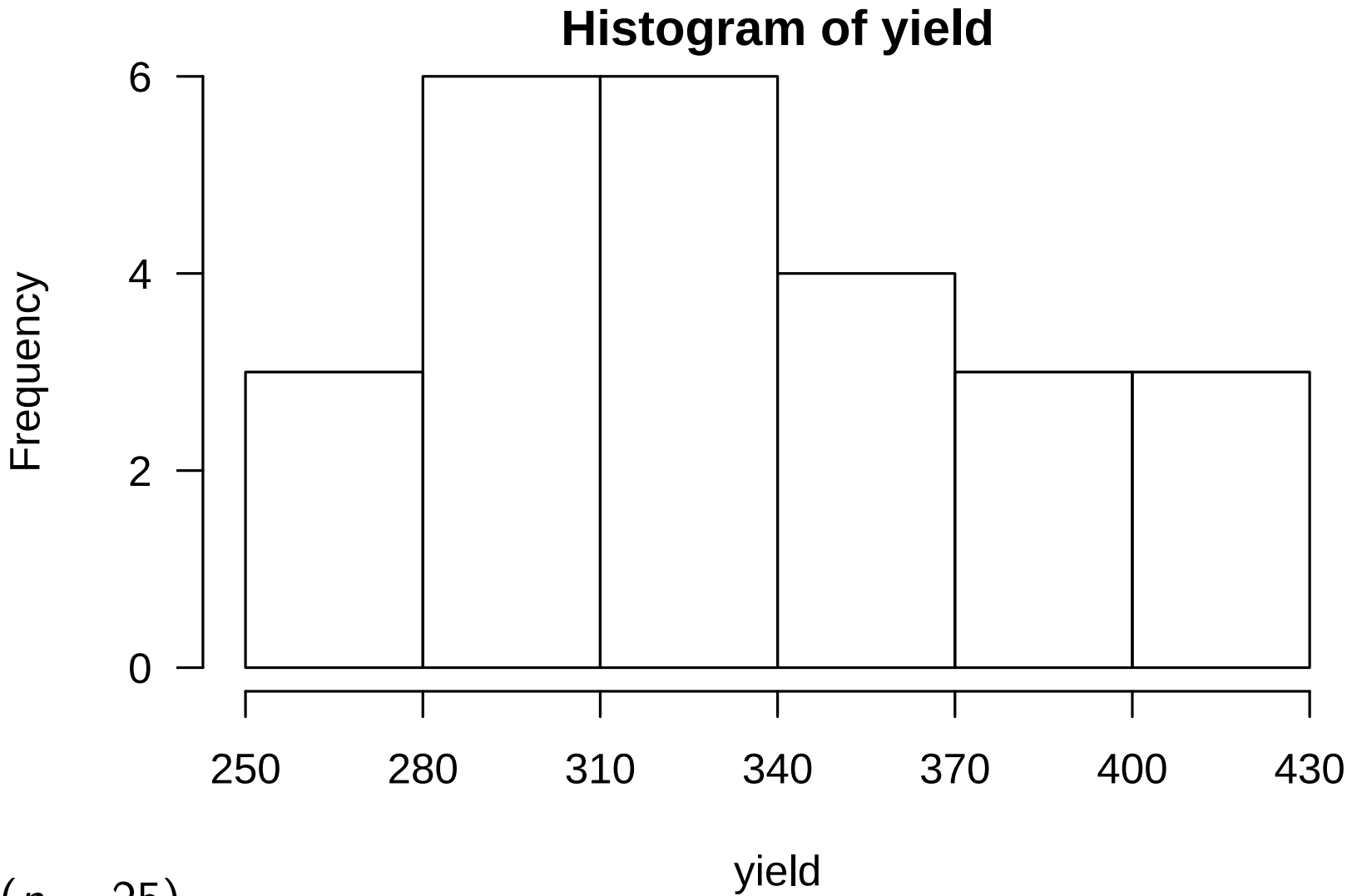
$$\begin{aligned} \text{yield} &= \text{mean} + \textit{error} \\ y_i &= \mu + e_i \\ e_i &\overset{\text{d}}{=} \text{N}(0, \sigma) \qquad i = 1, \ldots, 25. \end{aligned}$$

e_i represents separate draws from a $\text{N}(0, \sigma)$ distribution. Assumption of *constant variance*.

$\hat{\mu} = \bar{x}$; $\hat{\sigma} = s$; so, $\hat{\mu} = 333.3$; $\hat{\sigma} = 47.4$.

Statistical models – Example: Shape? Normal?

No explanatory variables – most basic model



$(n = 25)$

Statistical models – Example Lambs

One numerical explanatory variable

GROWTH OF LAMBS: All the lambs on a farm were weighed on a particular day. The age of each was known.

Age (days)	7	8	8	10	12	14	14	15	22	35	36	38
Weight (kg)	5.43	5.49	3.31	7.12	8.23	4.94	6.08	6.31	9.61	11.09	11.97	12.07

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
7	8	8	10	12	14	14	15	22	35	36	38
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
5.43	5.49	3.31	7.12	8.23	4.94	6.08	6.31	9.61	11.09	11.97	12.07

Statistical models – Example Lambs

One numerical explanatory variable

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
7	8	8	10	12	14	14	15	22	35	36	38
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
5.43	5.49	3.31	7.12	8.23	4.94	6.08	6.31	9.61	11.09	11.97	12.07

$$\text{lamb weight} = \alpha + \beta \times \text{age} + \text{error}$$

$$y_i = \alpha + \beta x_i + e_i$$

$$e_i \stackrel{\text{d}}{=} \text{N}(0, \sigma^2) \qquad i = 1, \ldots, 12$$

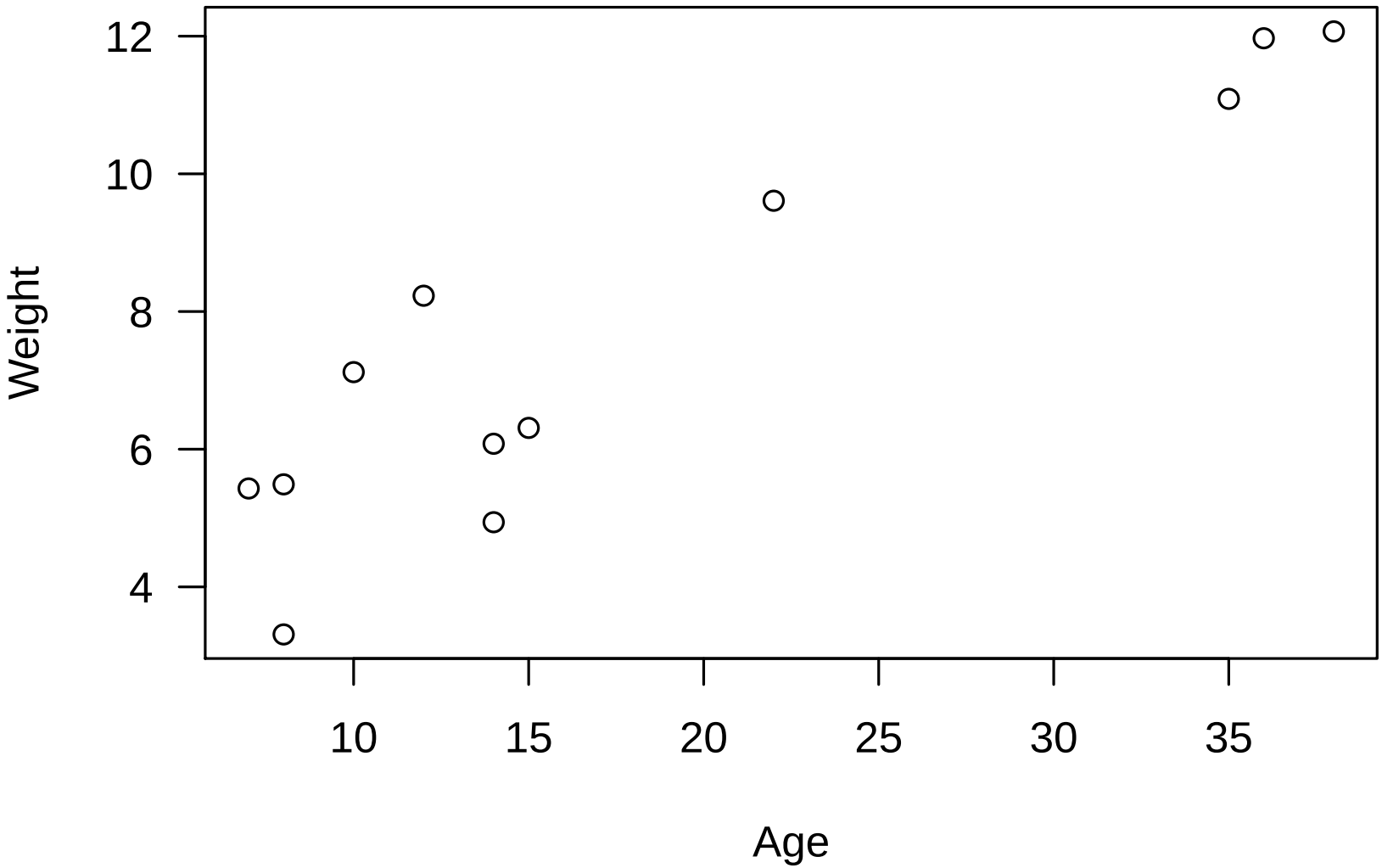
(check assumption for e_i)

$\hat{\alpha} =$ intercept of fitted line; $\hat{\beta} =$ slope of fitted line.

$\hat{\sigma}^2 = s^2 =$ residual mean square $= \frac{1}{n-2} \sum (\text{obs} - \text{fitted})^2$.

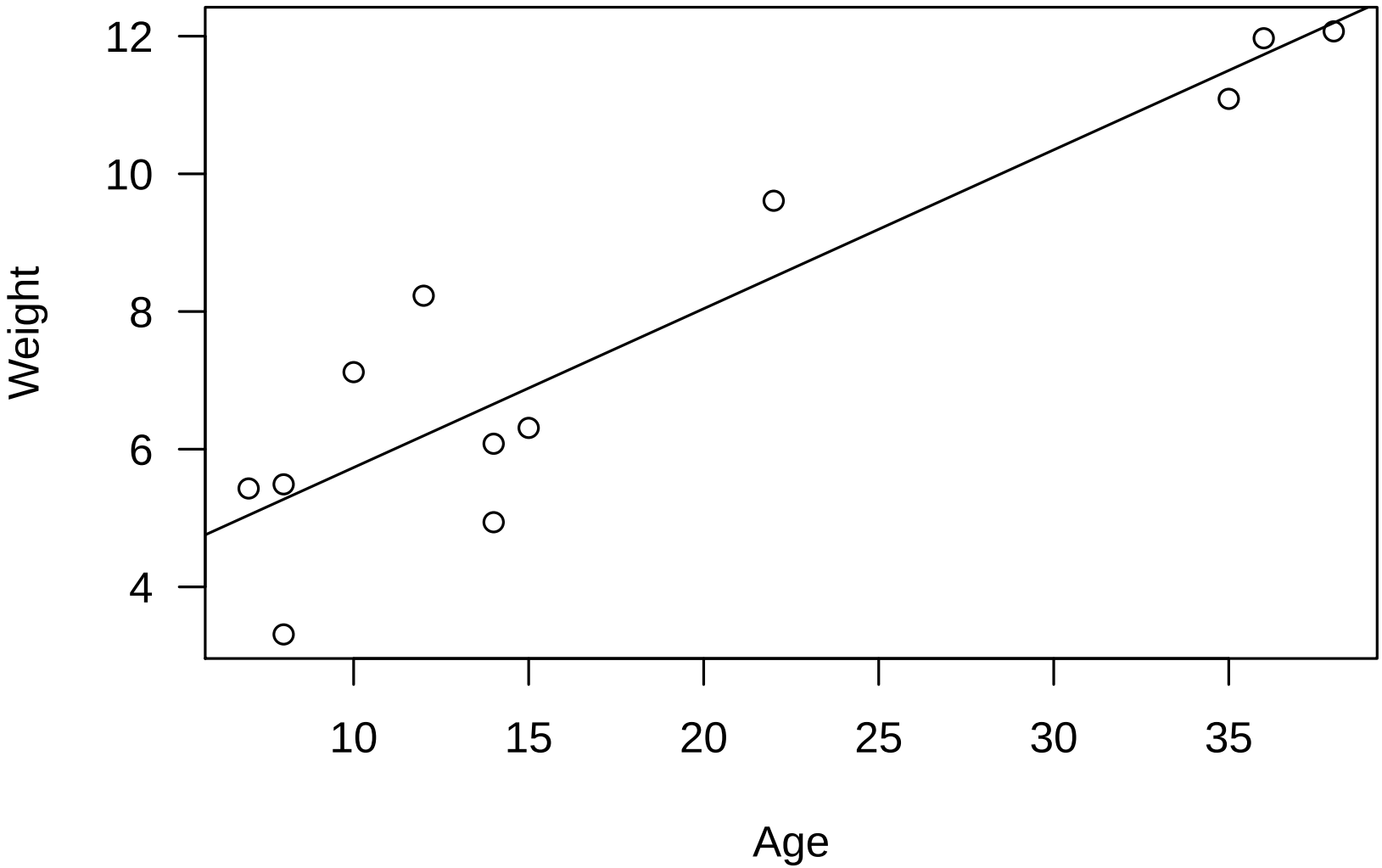
Statistical models – Example Lambs

One numerical explanatory variable



Statistical models – Example Lambs

One numerical explanatory variable



Statistical models – Example Bugs

One categorical explanatory variable (or factor)

COLOURS ATTRACTING BUGS: An experiment to examine how effective various colours were in attracting cereal leaf beetles to coloured boards in an oat field.

Colour	Beetles trapped					
Yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

$$\text{number of beetles} = \text{mean for colour used} + \text{error}$$

$$y_{ij} = \mu_i + e_{ij}$$

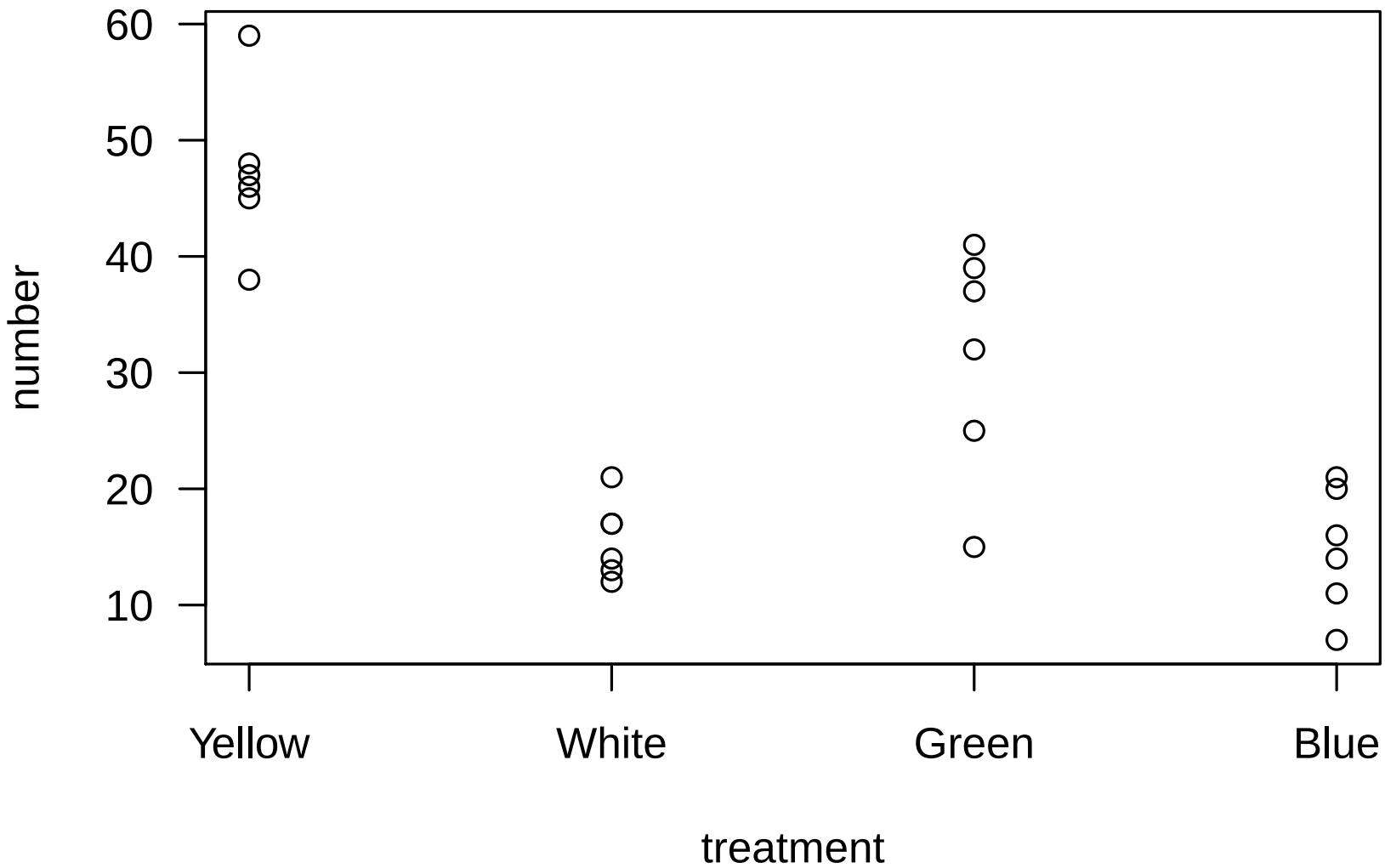
$$e_{ij} \overset{\text{d}}{=} \text{N}(0, \sigma^2) \qquad i{=}1, \ldots, 4; \quad j{=}1, \ldots, 6$$

$\hat{\mu}_1 = \bar{y}_1, \ldots;$

$\hat{\sigma}^2 = s^2 =$ residual mean square $= \frac{1}{n-k} \sum (\text{obs} - \text{fitted})^2.$

Statistical models

One categorical explanatory variable



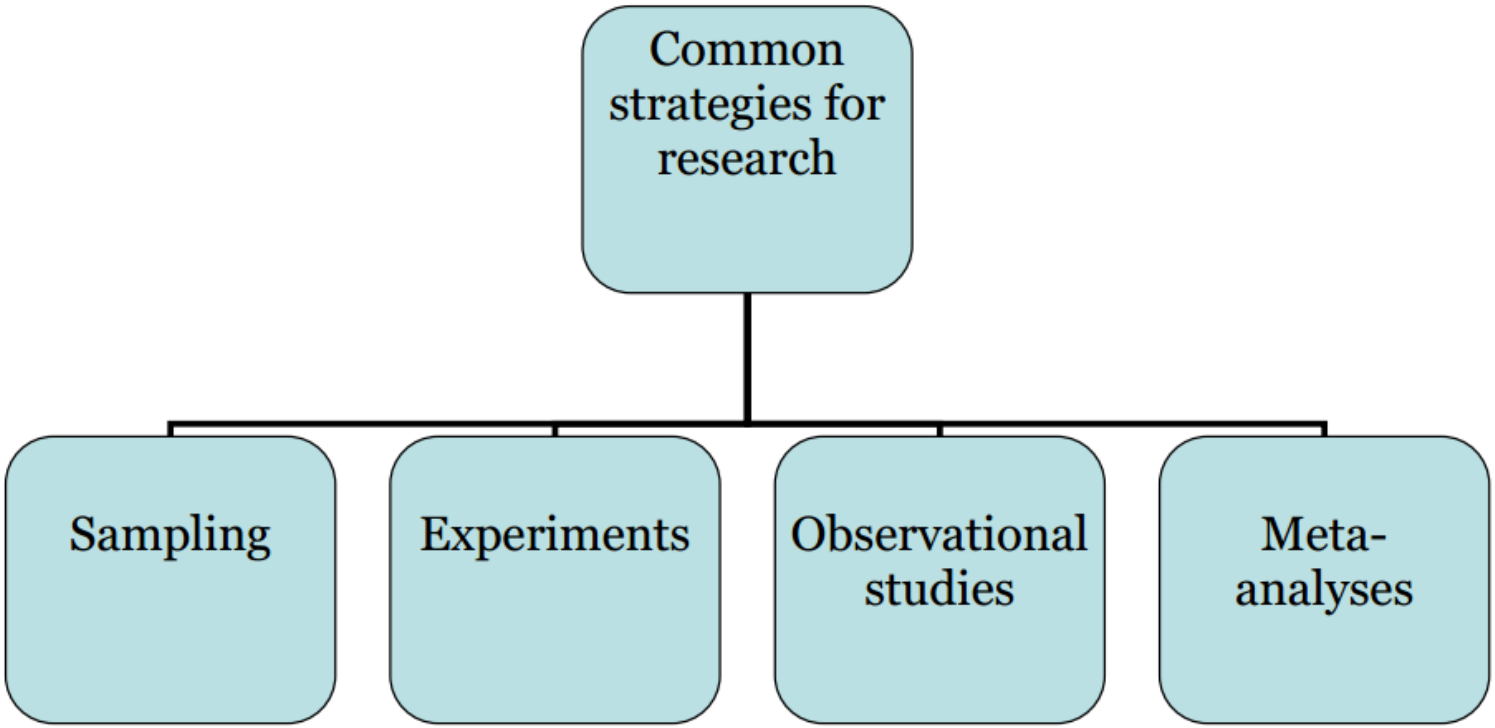
Outline

Statistical Models (model variation in the data, used to make inferences)

Sampling (how to select data)

Sampling Distributions (model sample-to-sample variation)

Sampling


Common strategies for research
Sampling
Experiments
Observational studies
Meta-analyses

Sampling

Simple random sampling

- From a population of units, each possible sample of size n is equally likely.
- Each unit in the population has the same probability of selection i.e. of appearing in the sample.
- Ensures that the estimates of population parameters are unbiased.

The sample will be *representative* of the population (?).
Note that a sampling *method* may be (un)biased, but not the particular sample itself.

Sampling

Random sampling

- random sample = iidrvs!
- model! (hyper-model?) random sampling is a process.
  It generates a representative sample (most likely)
  and one that we have a theoretical handle on.
  We use this as a 'template' for sampling. ($\overset{\circ}{\underset{\circ}{\bigcirc}}$)
- population/model $X \overset{\mathrm{d}}{=} \mathrm{N}(31, 5)$ or $X \overset{\mathrm{d}}{=} \mathrm{U}[1, 6]$

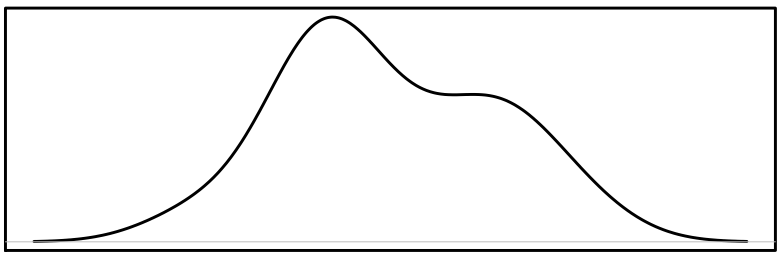
  $[X_1 \overset{\mathrm{d}}{=} \mathrm{N}(31, 5),\ X_2 \overset{\mathrm{d}}{=} \mathrm{N}(31, 5),\ \ldots,\ X_{10} \overset{\mathrm{d}}{=} \mathrm{N}(31, 5)]$

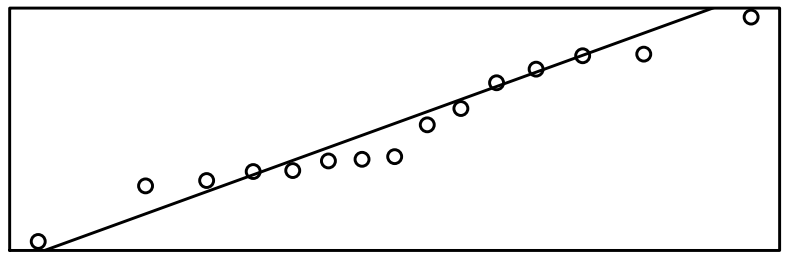
  In the models above $e_i \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma)$ ... iidrvs!

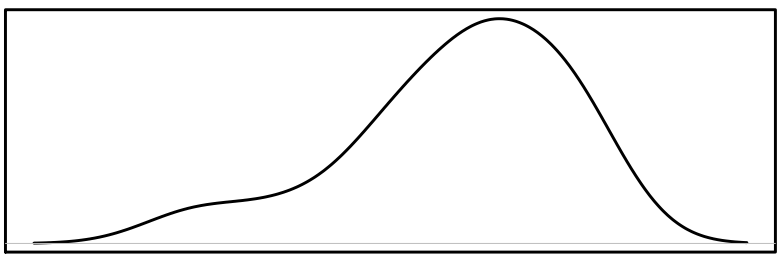
Sampling from a Normal Distribution: $n = 15$

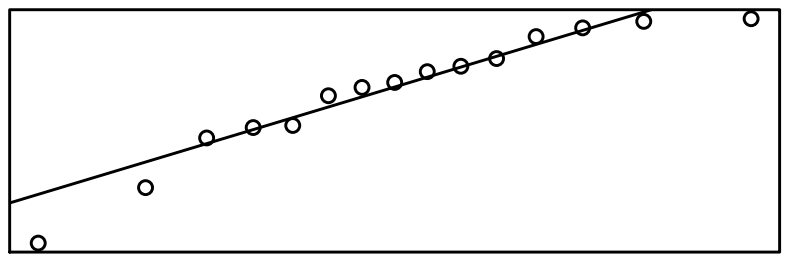
density()
(smoothed histogram)

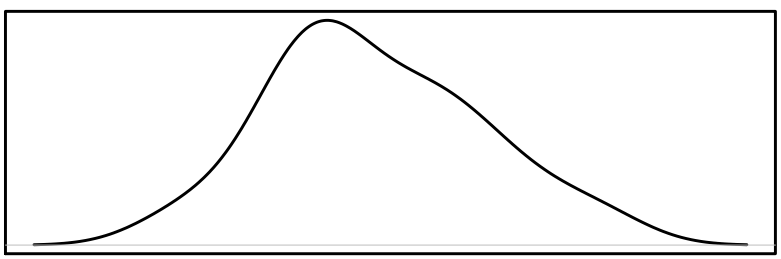
qqplot()
(accumulated data vs
theoretical normal quantiles)

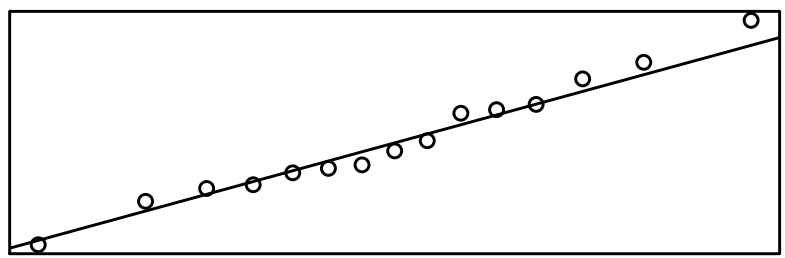




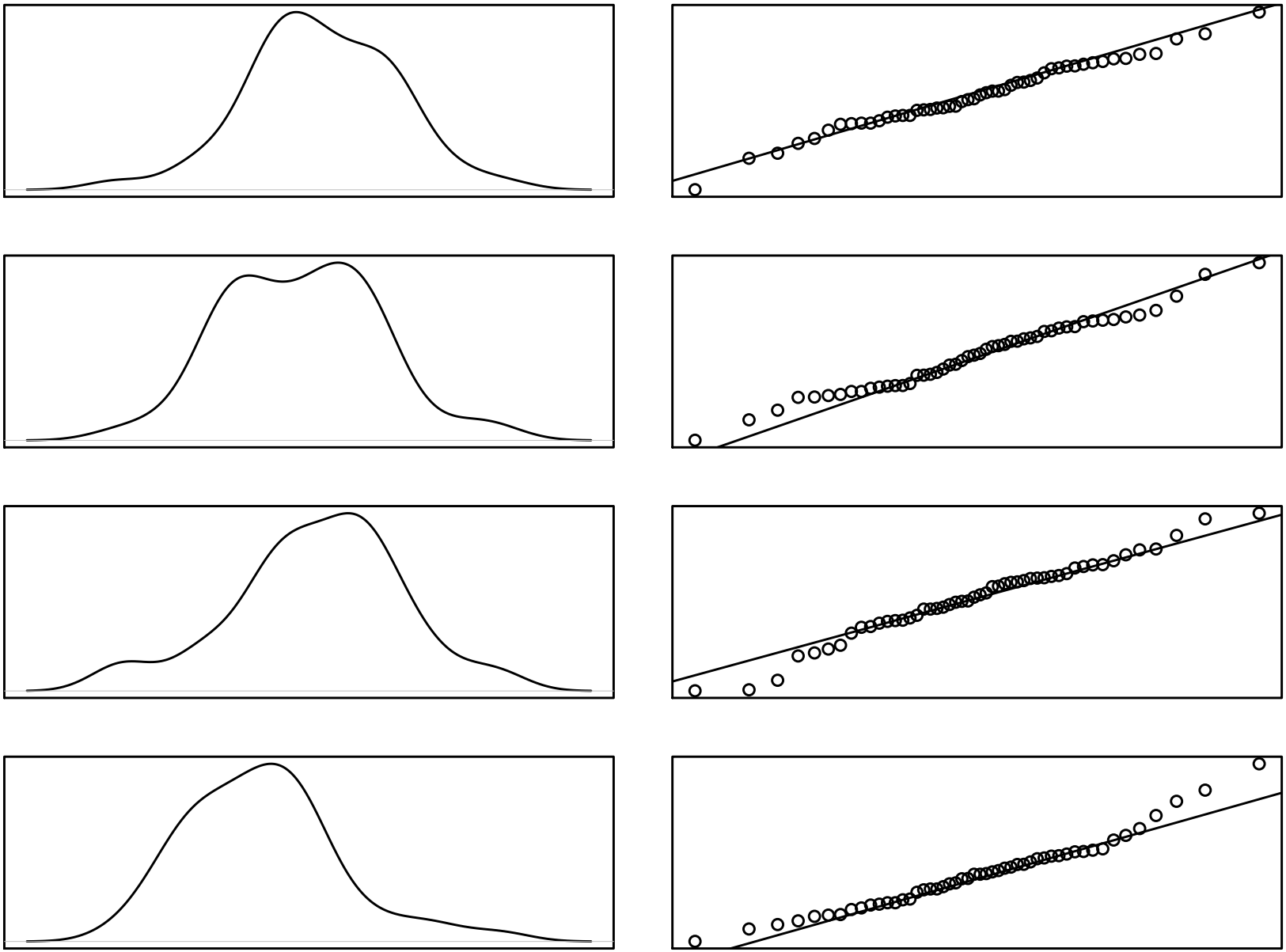




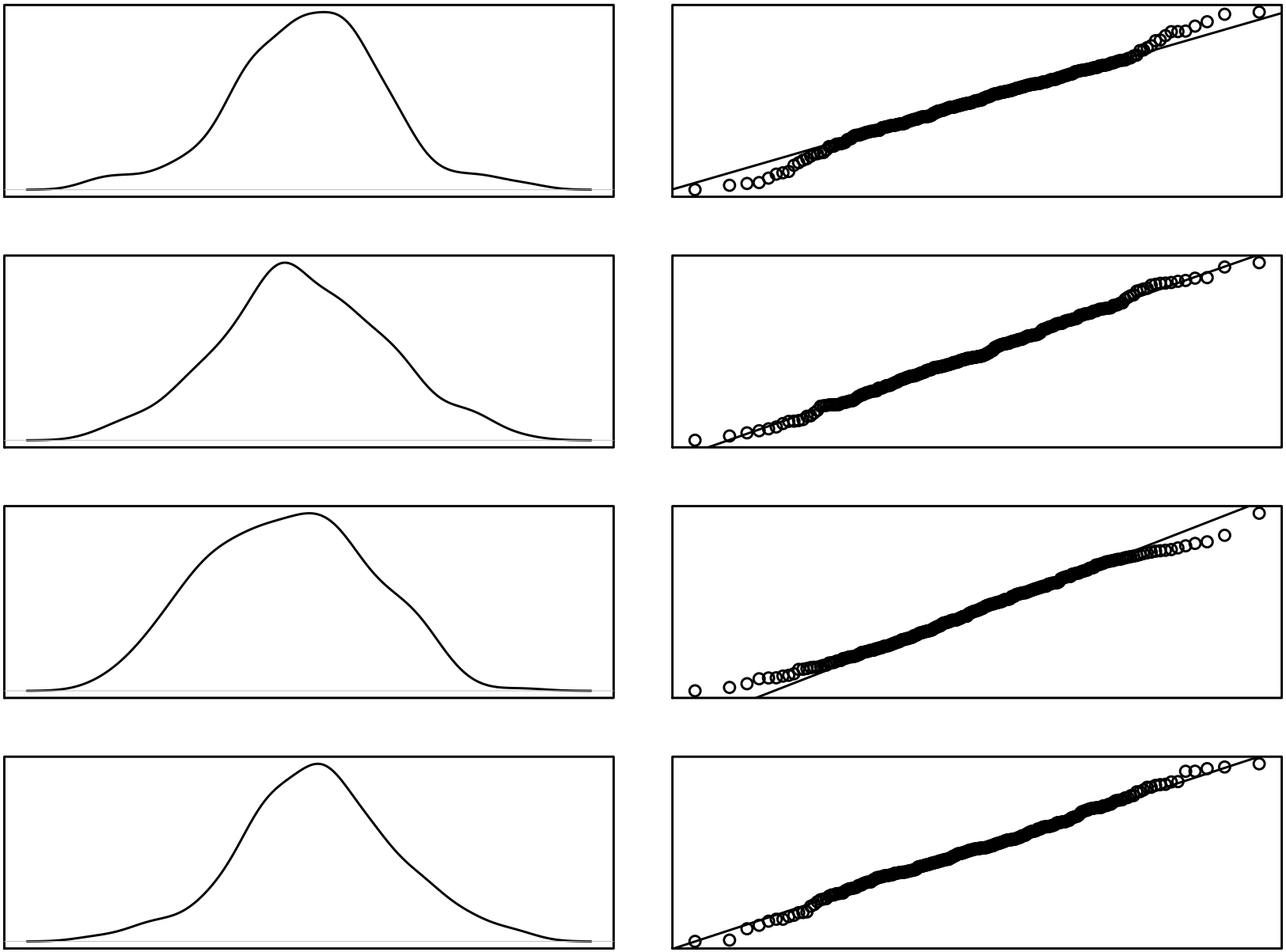




Sampling from a Normal Distribution: $n = 50$



Sampling from a Normal Distribution: $n = 300$



Sampling Distributions

We can examine the sampling distribution of any statistic calculated from a sample.

The sample mean is usually the most important.

Examples: blood pressure (mm Hg) of adult males; diameter (cm) of mass-produced bolts; yield (kg/hA) from a rye crop; length (mm) of beetles; ...

Could take repeated samples of size $n = 10$, and calculate the sample mean each time. ( )

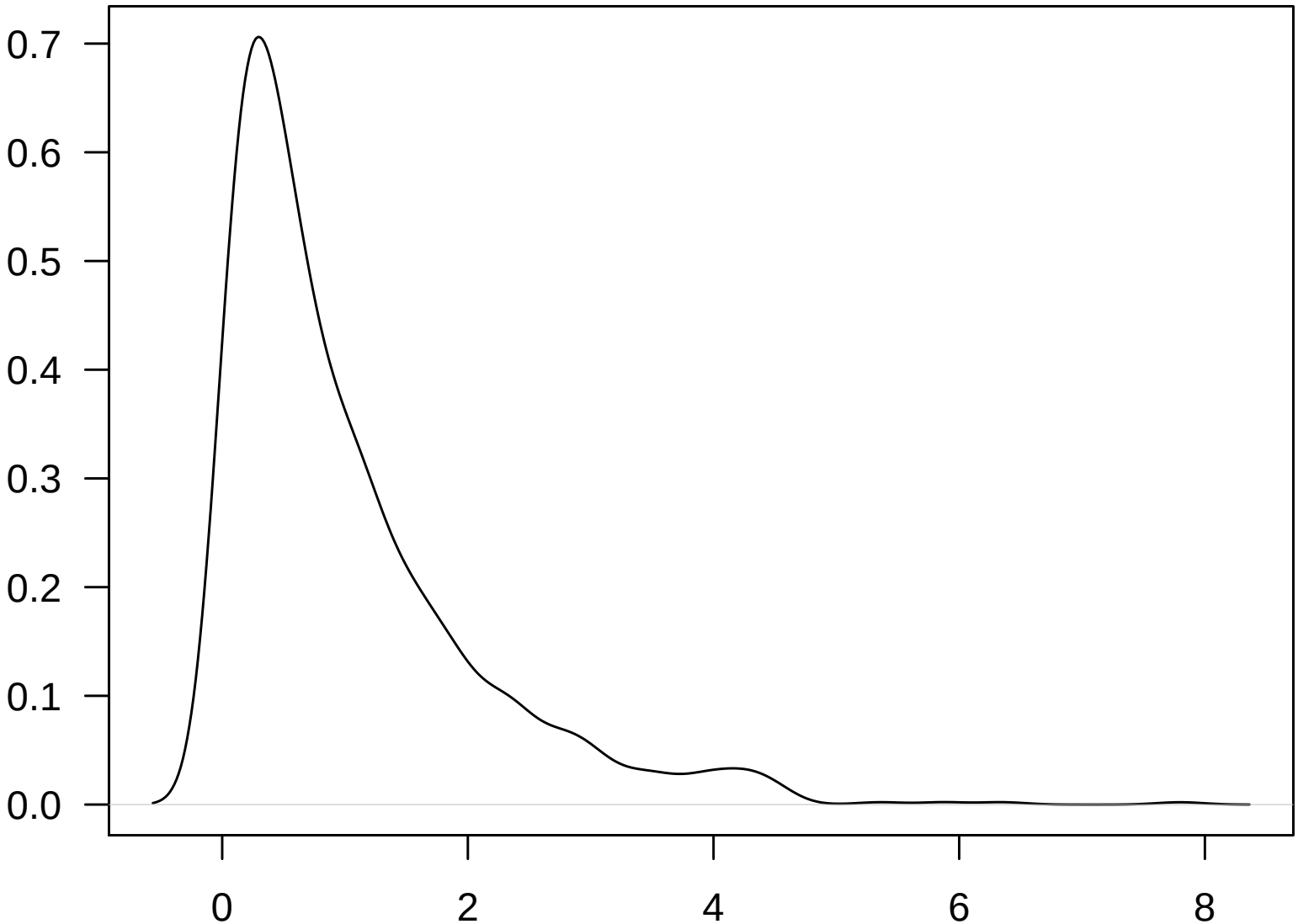
The variation will be less than that of the individual observations.

How much less? The theoretical underpinnings of random sampling means we can work this out.

Sampling Distributions

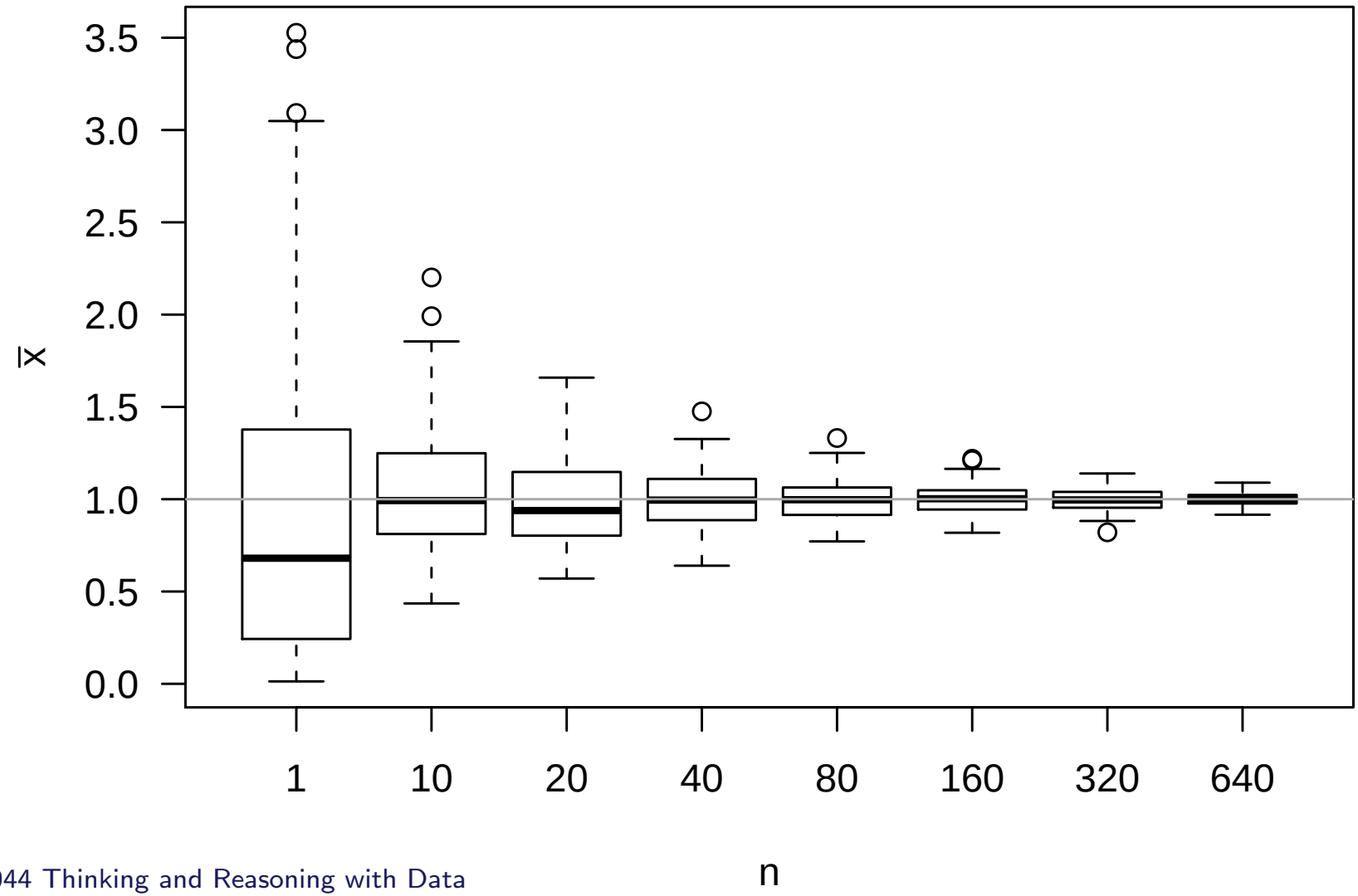
$$\mathrm{E}(\bar{X}) = \mu, \qquad \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

Sampling from an exponential distribution: $n = 1000$



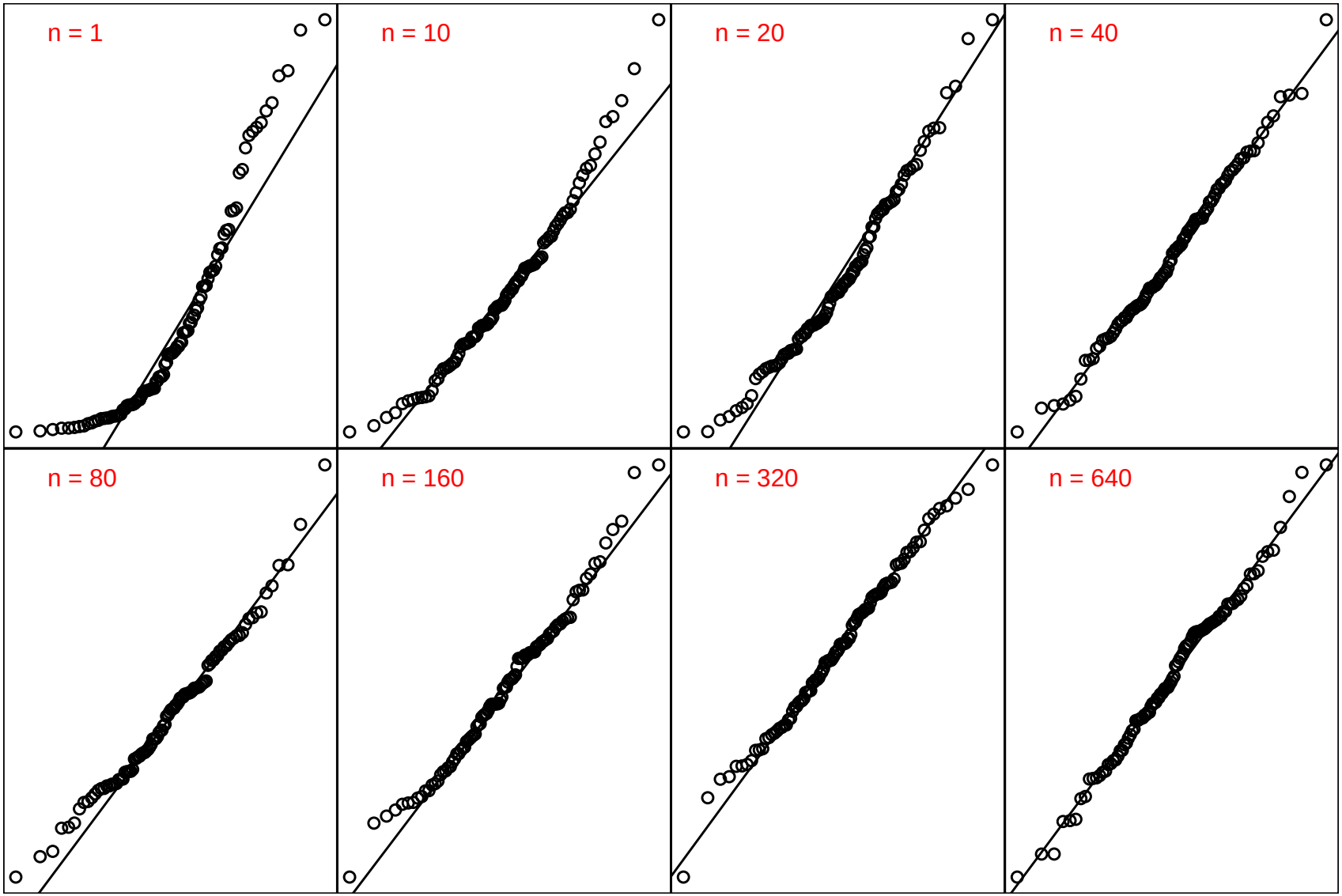
Sampling distribution of the sample mean

Boxplots: Exponential distribution with rate $= 1$: $X \overset{\mathrm{d}}{=} \exp(1)$
100 samples for each n



Sampling distribution of the sample mean

qqplots: Exponential distribution with rate $= 1$: $X \overset{\mathrm{d}}{=} \exp(1)$



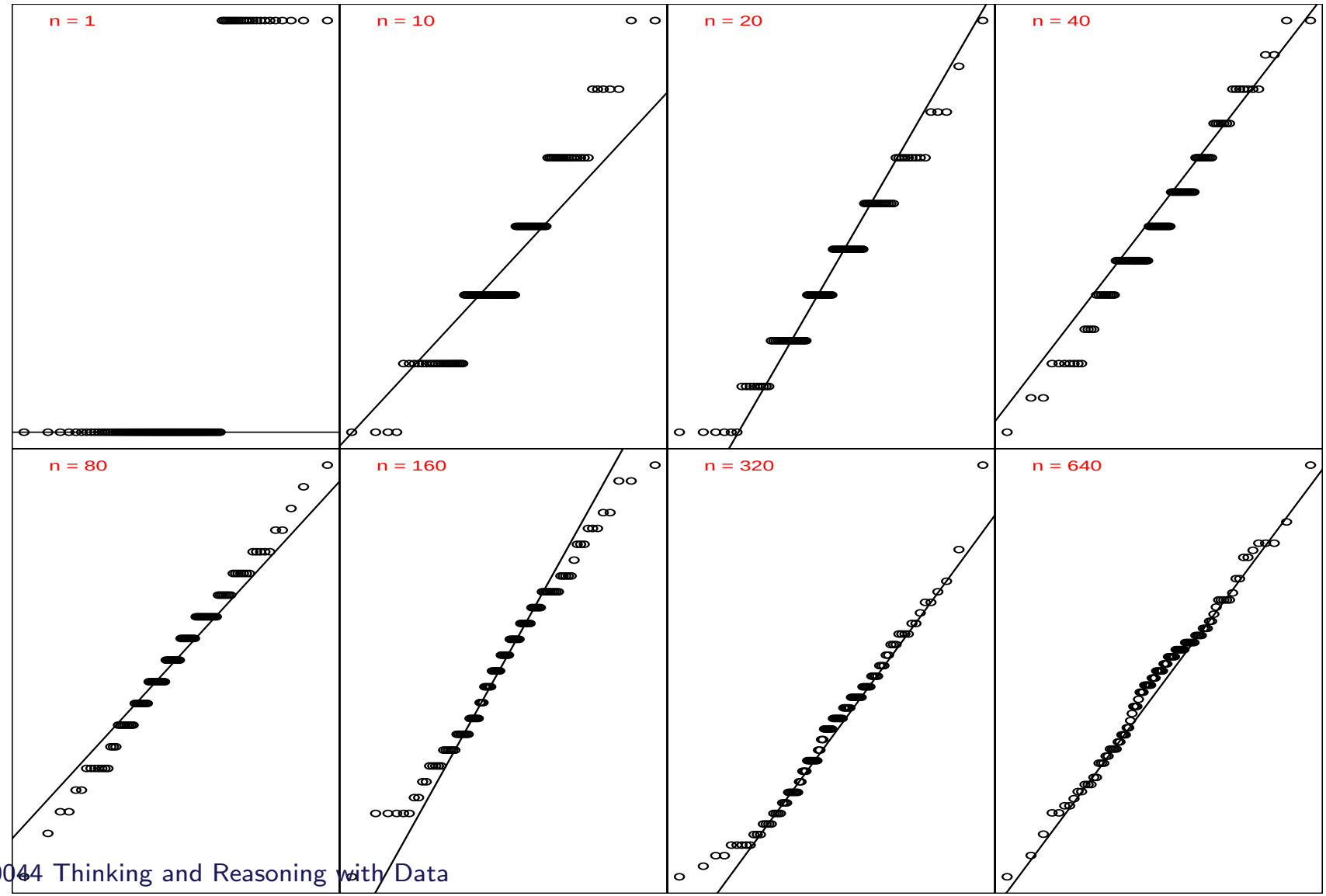
The central limit theorem

For a sufficiently large number of independent random variables, the sampling distribution of the sample mean is approximately normal, *no matter what the distribution of the random variable*.

Consequence: many statistical techniques are robust to non-normality, especially those that involve sums.

Sampling distribution of the sample mean

qqplots: Binomial distribution with $p = 0.25 : X \overset{\mathrm{d}}{=} \mathrm{Bi}(n, 0.25)$



The central limit theorem

For a sufficiently large number of independent random variables, the sampling distribution of the sample mean is approximately normal, *no matter what the distribution of the random variable*.

Consequence: many statistical techniques are robust to non-normality, especially those that involve sums.

The central limit theorem

Central limit theorem (CLT)

The sum of a large number of identically distributed random variables which are independent is approximately normally distributed. This holds no matter the distribution of the individual variables.

If $T = X_1 + X_2 + \cdots + X_n$, where $X_1, X2, \ldots, Xn$ are independent observations on X, and $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$, then we already know that $\mathrm{E}(T) = n\mu$, $\mathrm{var}(T) = n\sigma^2$, $\mathrm{sd}(T) = \sigma\sqrt{n}$. The CLT states that, if n is large, $T \approx N(n\mu, n\sigma^2)$.

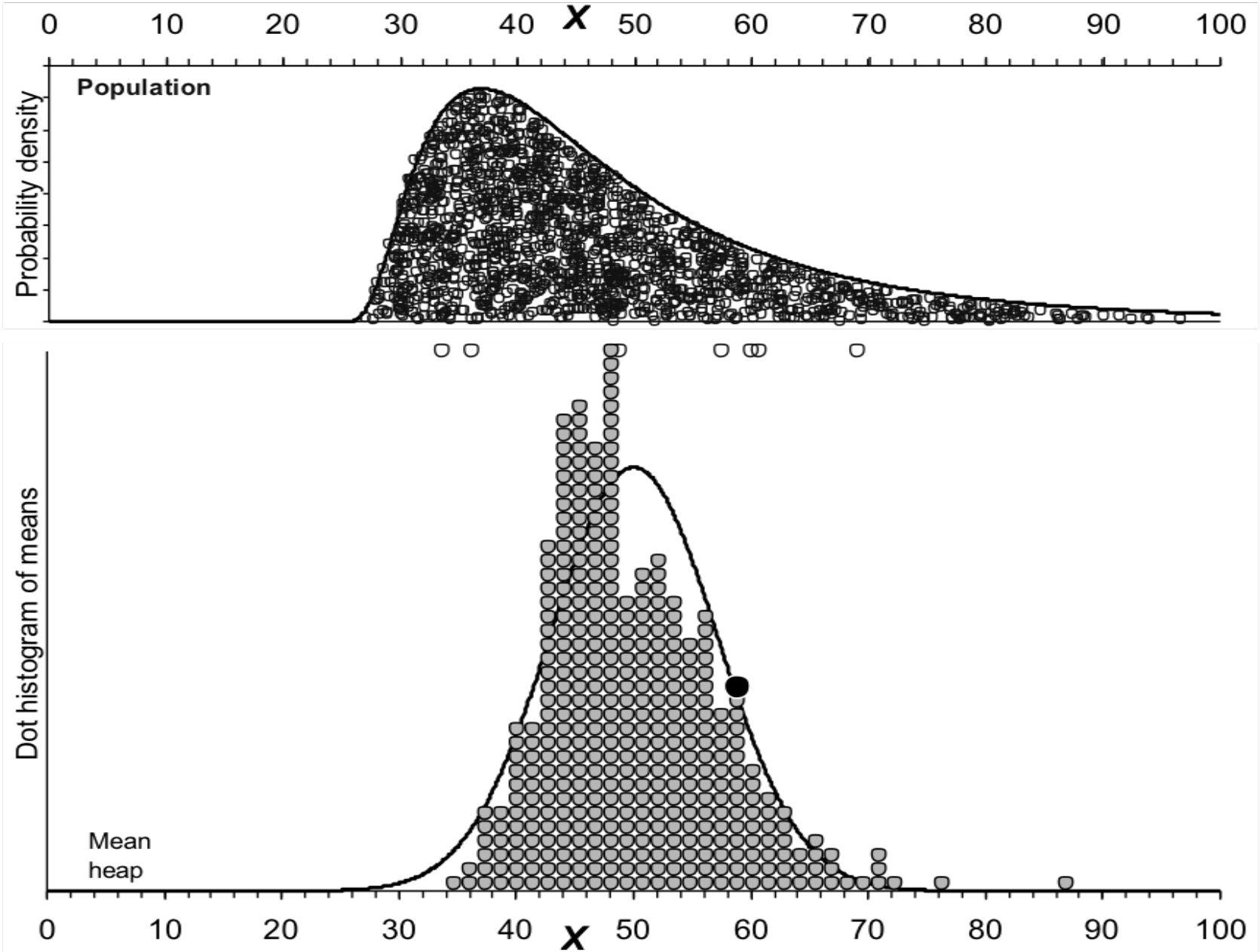
CLT and random sampling

Since $\bar{X} = \frac{1}{n} \times$sum, the Central Limit Theorem says that:

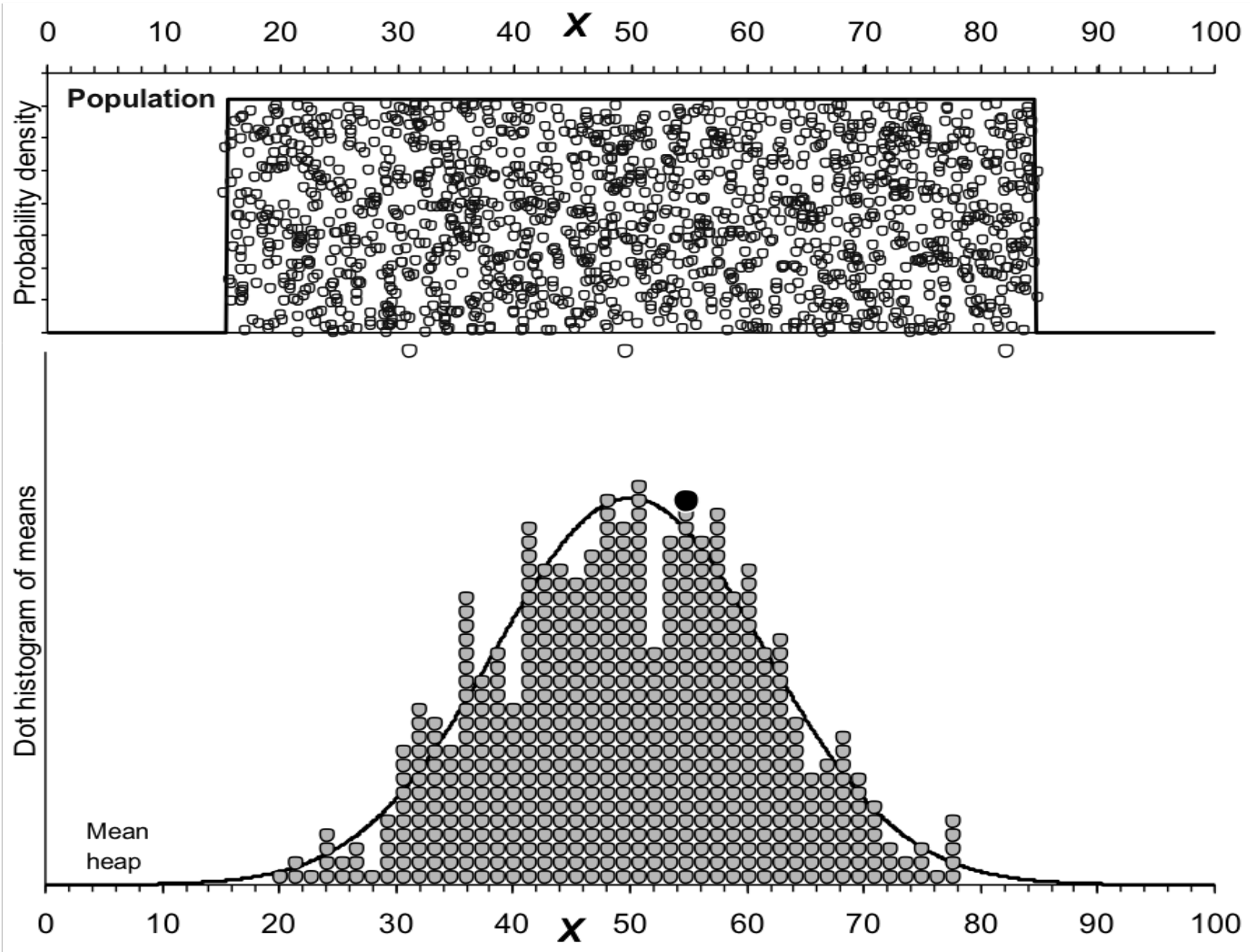
$$\bar{X} \overset{\mathrm{d}}{\approx} N(\mu, \frac{\sigma^2}{n}).$$

The sampling distribution of the sample mean will follow approximately a Normal distribution.

CLT and Random Sampling



CLT and Random Sampling



Order statistics and quantiles

If $x_1, \ldots, x_n$ are distinct numbers then we can order them uniquely from smallest to largest as

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}.$$

We may then be interested in the "quantiles" of this set of numbers. For $1 \leqslant k \leqslant n$ we want to associate to $x_{(k)}$ a "quantile" $q(k) \in [0, 1]$, but what is the right choice of $q(k)$?

There are many different versions of this, but we will use $q(k) = k/(n+1)$. In other words we will associate to the k-th number in the set the quantile $k/(n+1)$.

Example: If the ordered values are $1.1, 1.6, 1.8, 3.0, 5.1$ then the quantile associated to the value 1.8 is the $3/(5+1) = 1/2$ quantile (i.e. the median).

Order statistics and quantiles

We can apply the above to a random sample.

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution.
- The ordered values $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics**. In particular $X_{(1)}$ is the minimum of the sample and $X_{(n)}$ is the maximum.
- From the above discussion it is natural to estimate the true $q(k)$-quantile of the unknown distribution as $X_{(k)}$. Thus, if $n = 5$ then (with $q(k) = k/(n+1)$ as above) we would estimate the median (which is the $q = 1/2$ quantile) of the unknown distribution by $X_{(3)}$, where $1/2 = 3/(5+1)$.

Order statistics and quantiles

More generally, a natural estimator for the q quantile is written $\hat{c}_q = X_{(k)}$, where $k = (n+1)q$. If this does not give an integer then you have a choice to make. Many people interpolate between the two values on either side.

Example: If $n = 5$ and we are interested in the .75 quantile, then $k = 6 * .75 = 4.5$ so we could estimate the .75 quantile by $(X_{(4)} + X_{(5)})/2$.

These estimators are called the **sample quantiles**.

Order statistics and quantiles – Example

So, if we generated a random sample $X_{(1)}, \ldots, X_{(n)}$ from a normal distribution with mean ν and variance σ^2, and we plot the sample quantiles of the observed values $x_1, \ldots, x_n$ on the y-axis and the quantiles of the standard normal distribution on the x axis, we would expect to see points that are fairly close to a straight line with intercept μ and slope σ (but not exactly on a straight line, because they are random). If n is really large we might expect the points to be very close to a straight line.

Check for Normality — the qq-plot

If $X \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \stackrel{\mathrm{d}}{=} \mathrm{N}(0,1) \quad \Rightarrow \quad X \stackrel{\mathrm{d}}{=} \mu + \sigma Z.$

$$\underset{\text{the } q\text{-quantile for } X}{c_q(X)} = \mu + \sigma z_q$$

Now, $x_{(k)} \sim c_q(X)$, where $q = \frac{k}{n+1} \quad \Rightarrow \quad x_{(k)} \sim \mu + \sigma z_{k/(n+1)}$

SO ... if we plot the order statistics on the y-axis and the standard normal quantiles on the x-axis, then we should get a straight line with intercept μ and slope σ ... PROVIDED $X \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$.

This is effectively plotting the sample quantiles (order statistics) against the standard population quantiles — hence a qq-plot.

Checking Normality

How can we tell if a sample is normal (i.e. from a normal population)?

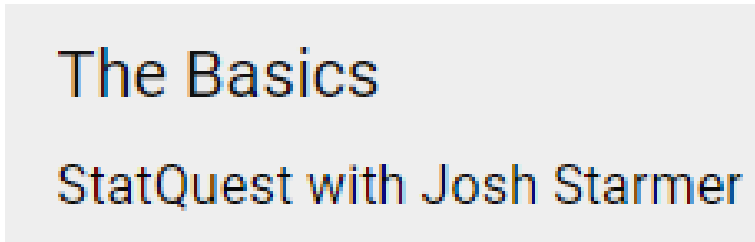
The sample pdf is too erratic to be much use.

The sample cdf is a bit more stable.

But how do we know which shape is a normal cdf?

Principle: The easiest curve to fit is a straight line

Let's watch ...



- https://www.youtube.com/watch?v=okjYjCISjOg

Example

Our definition of sample quantiles suggests a solution:

$$\hat{c}_q = x_{(k)}, \quad \text{where } k = (n+1)q.$$

Therefore:

$$x_{(k)} \sim c_q \quad \text{where } q = \frac{k}{n+1}.$$

We have discussed this in relation to the five-number summary for a sample from a Normal distribution.

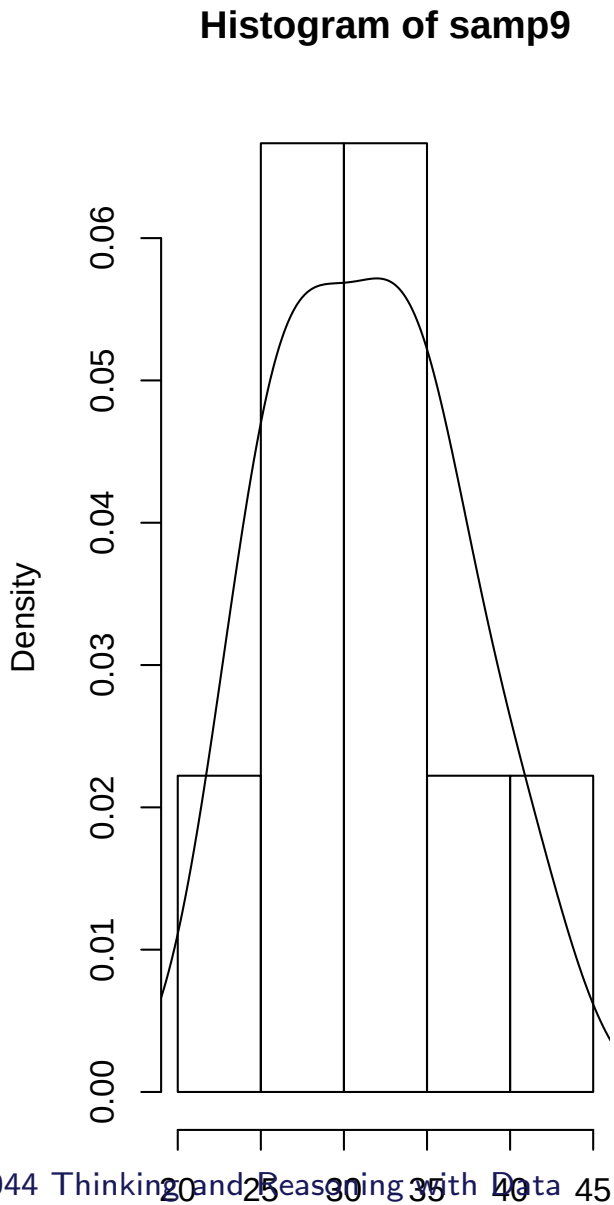
▷ **Example.** Let's look at a sample of $n = 9$

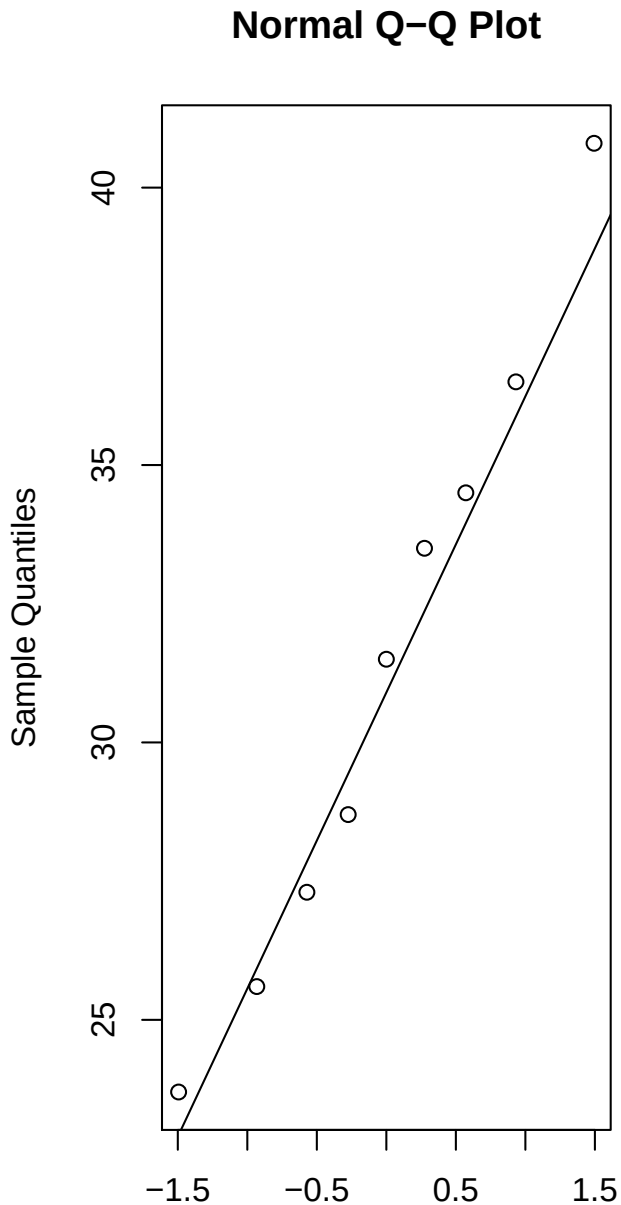
(because the numbers are easy: $q = \frac{k}{10} \Rightarrow 0.1, 0.2, \ldots, 0.9$)

34.5, 23.7, 28.7, 40.8, 33.5, 25.6, 36.5, 31.5, 27.3.

Example random sample of 9 observations

int: $\hat{\mu} = 31.3$; slope: $\hat{\sigma} = 6.7$ versus $\bar{x} = 31.3$ and $s = 5.5$





Checking Normality

40.8	$x_{(9)}$	$\sim$	$c_{0.9}$	$=$	$\mu + 1.2816\sigma$
36.5	$x_{(8)}$	$\sim$	$c_{0.8}$	$=$	$\mu + 0.8416\sigma$
34.5	$x_{(7)}$	$\sim$	$c_{0.7}$	$=$	$\mu + 0.5244\sigma$
33.5	$x_{(6)}$	$\sim$	$c_{0.6}$	$=$	$\mu + 0.2533\sigma$
31.5	$x_{(5)}$	$\sim$	$c_{0.5}$	$=$	$\mu + 0.0000\sigma$
28.7	$x_{(4)}$	$\sim$	$c_{0.4}$	$=$	$\mu - 0.2533\sigma$
27.3	$x_{(3)}$	$\sim$	$c_{0.3}$	$=$	$\mu - 0.5244\sigma$
25.6	$x_{(2)}$	$\sim$	$c_{0.2}$	$=$	$\mu - 0.8416\sigma$
23.7	$x_{(1)}$	$\sim$	$c_{0.1}$	$=$	$\mu - 1.2816\sigma$
$\uparrow$					$\uparrow$
"y"					"x"

"y" $\sim \mu + \sigma$ "x"

order statistics

normal scores

sample **Q**uantiles

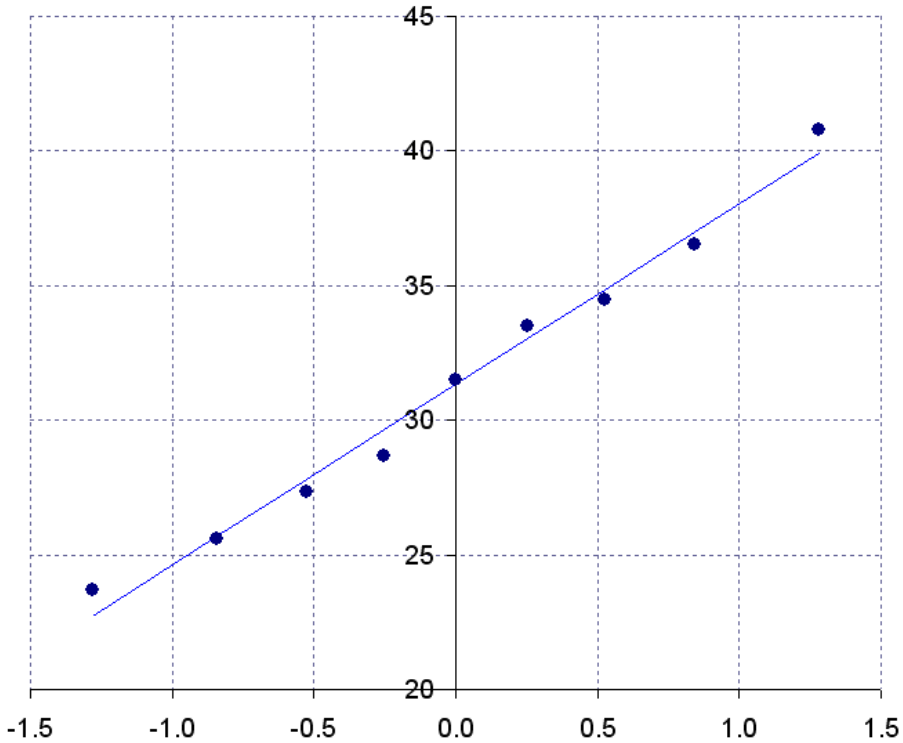
normal **Q**uantiles

QQ plot

$x_{(k)}$

$\Phi^{-1}(\frac{k}{n+1})$

Checking Normality



This appears to be reasonably close to a straight line, so the normal distribution is a reasonable model for these data.

The intercept of the fitted line gives an estimate of μ: $\hat{\mu} = 31.3$;
and the slope of the fitted line gives an estimate of σ: $\hat{\sigma} = 6.7$.

The quantities $\Phi^{-1}(\frac{k}{n+1})$ are called *normal scores*.

(values you would 'expect' for a sample from a standard normal distribution)

R actually uses modified (improved) normal scores: $\Phi^{-1}(\frac{k-\frac{3}{8}}{n+\frac{1}{4}})$, which are a bit more spread. R: `qqnorm(x)`

This plot is called a QQ-plot because it plots the sample Quantiles against the (standard) population Quantiles.

The quantities $\Phi^{-1}(\frac{k}{n+1})$ are called *normal scores*.

(values you would 'expect' for a sample from a standard normal distribution)

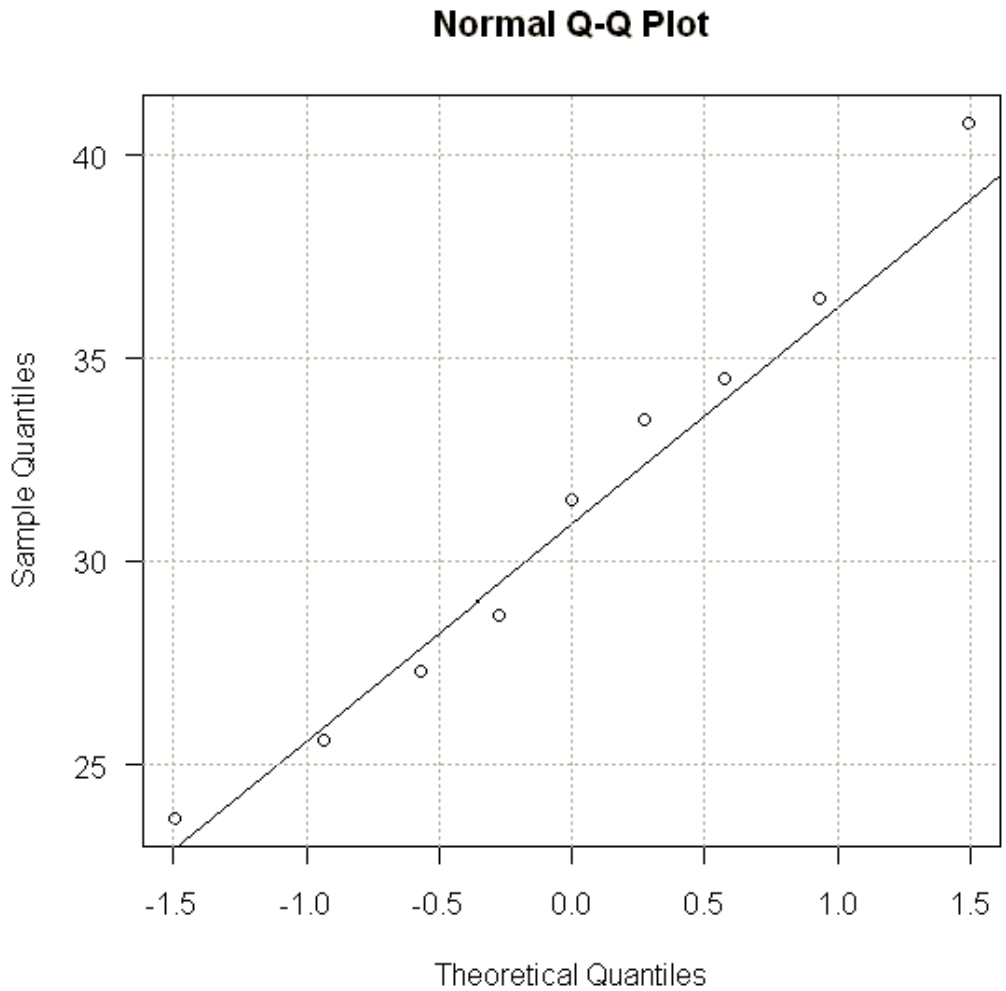
R actually uses modified (improved) normal scores: $\Phi^{-1}(\frac{k-\frac{3}{8}}{n+\frac{1}{4}})$, which are a bit more spread. R: `qqnorm(x)`

This plot is called a QQ-plot because it plots the sample Quantiles against the (standard) population Quantiles.

The QQ-plot not only provides an indication of whether the model is a reasonable fit, but also gives estimates of μ and σ (the intercept and slope of the fitted line).

These estimates work even in some situations where $\bar{x}$ and s won't. For example censored or truncated data.

```
> qqnorm(mydata,las=1)
> qqline(mydata)
> grid(col="darkgray")
```



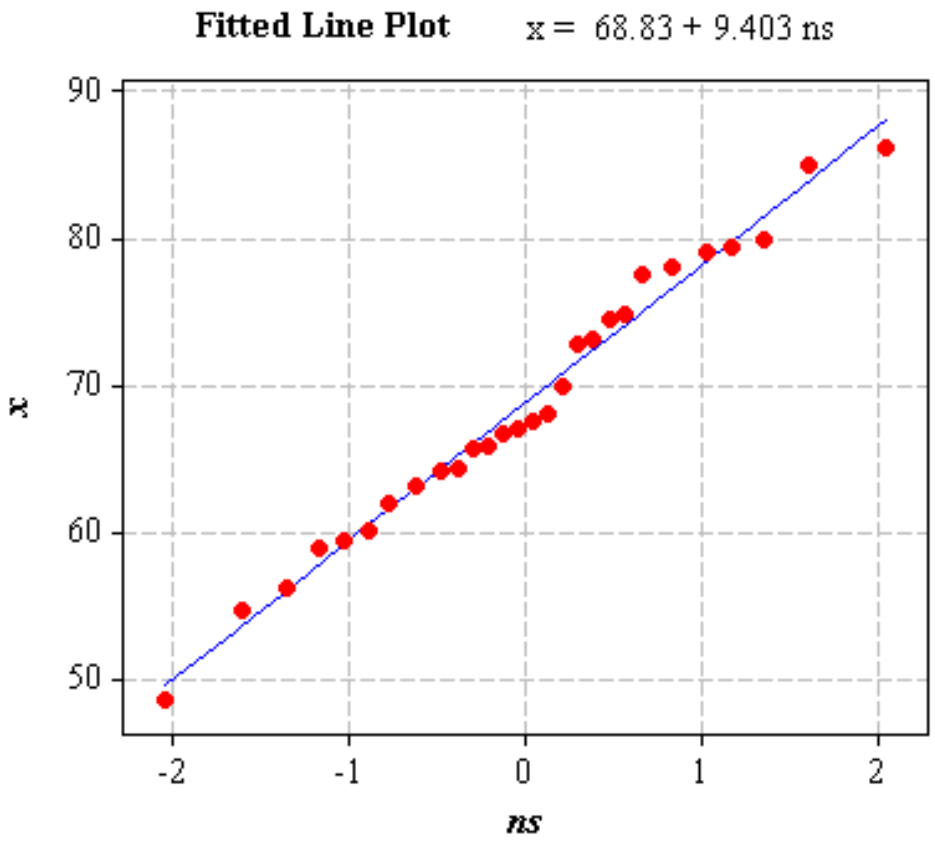
▷ **Example.** A random sample of 30 observations gives:

68.1 73.1 86.2 85.1 70.0 67.1 64.3 65.8 64.2 62.0 $\cdots$ 79.5 67.6

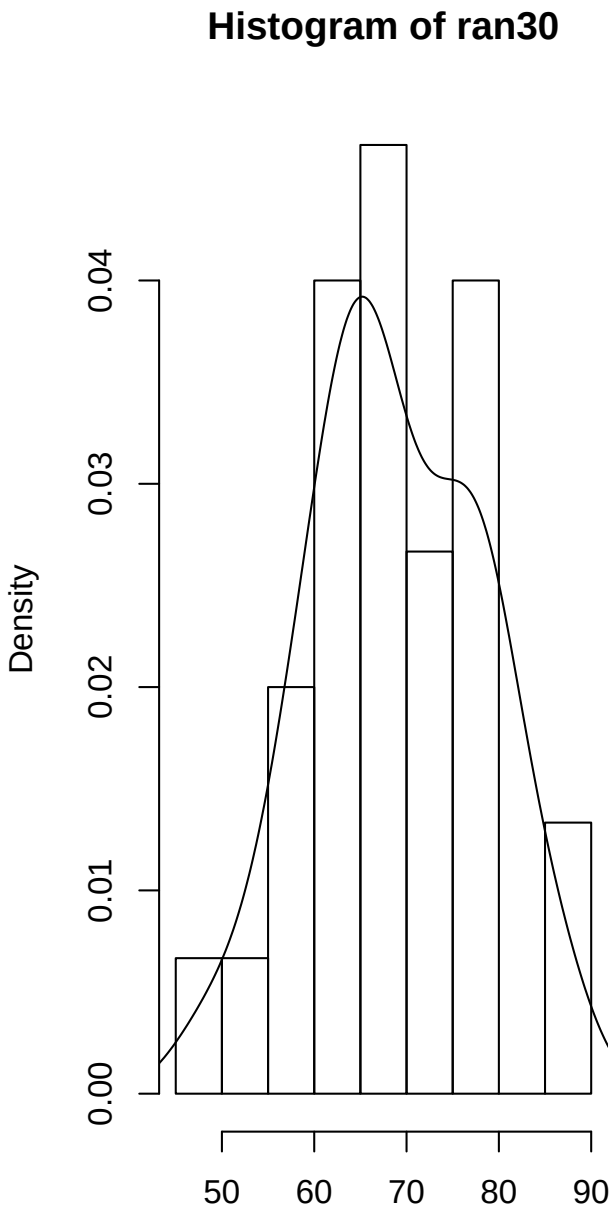
Is this a sample from a normal population?

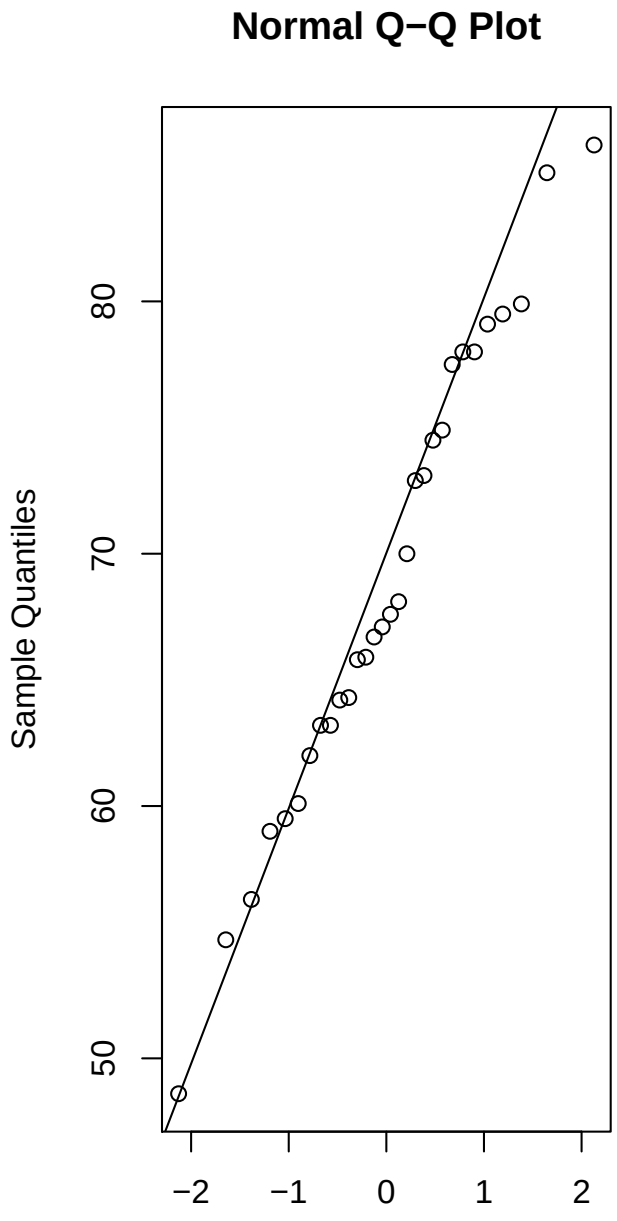
The dotplot/histogram don't show much.

The QQplot gives a more useful guide.

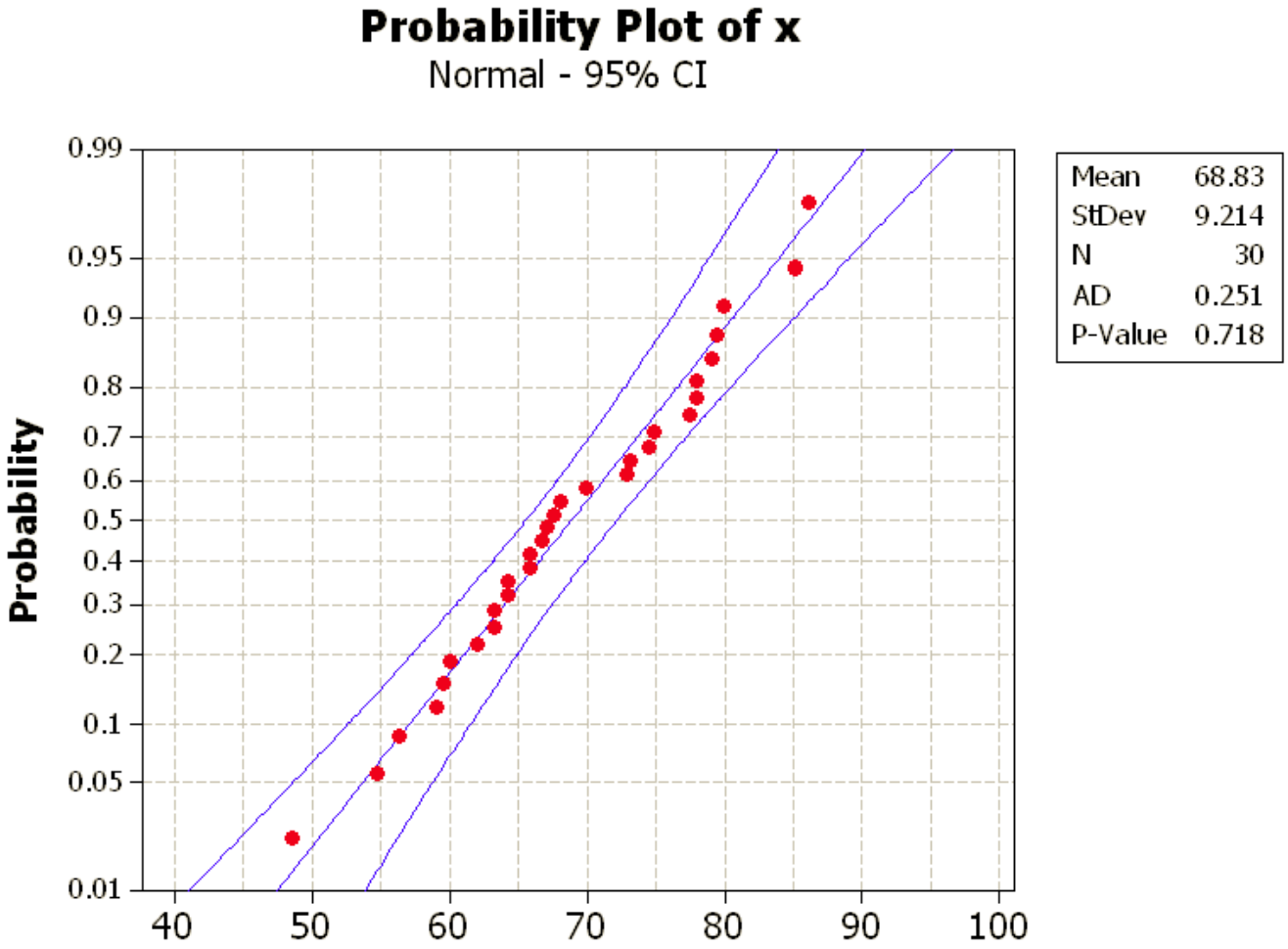


Example random sample of 30 observations

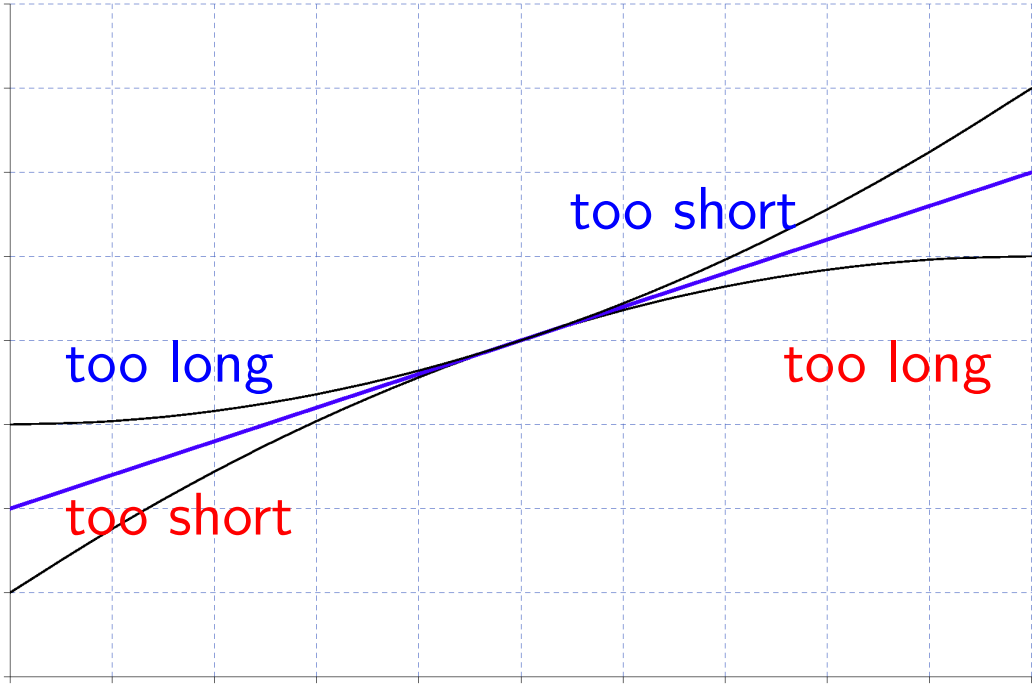




With the axes reversed, it is essentially a plot of the cdf on a warped scale. This is called a Probability plot and is what some packages do routinely:



What happens if it's a bad fit? *tails too long; or tails too short.*



A concave graph indicates too short at one end and too long at the other, i.e. a skew distribution.

Example random sample from Exp(1), $n = 200$