AIMS VIDEO COURSES

SUPPORTING BOOKLET

PROBABILITY & STATISTICS

WITH

PROF DAVID SPIEGELHALTER

AIMS

SOUTH AFRICA

AIMS
SOUTH AFRICA

African Institute for Mathematical Sciences

6 MELROSE ROAD | MUIZENBERG | CAPE TOWN 7945 | SOUTH AFRICA
TEL: +27 (0)21 787 9320 | FAX: +27 (0)21 787 9321
EMAIL: info@aims.ac.za | WEB: www.aims.ac.za

AIMS Online Courses

The mission of the AIMS academic programme is to provide an excellent, advanced education in the mathematical sciences to talented African students in order to develop independent thinkers, researchers and problem solvers who will contribute to Africa's scientific development.

Teaching at AIMS is based on the principle of learning and understanding, rather than simply listening and writing, during classes, and on creating an atmosphere of increasing our knowledge through class discussions, through small group discussions, by formulating conjectures and assessing the evidence for them, and sometimes going down wrong paths and learning from the mistakes that led us there. The essential features of the classes at AIMS are that, in contrast to formal lecture courses, they are highly interactive, where the students engage with the lecturer throughout the class time, are encouraged to learn together in a journey of questioning and discovery, and where lecturers respond to the needs of the class rather than to a pre-determined syllabus. AIMS teaching philosophy is to promote critical and creative thinking, to experience the excitement of learning from true understanding, and to avoid rote learning directed only towards assessment.

Leading international and local experts offer the courses at AIMS, which are three weeks long (each module consisting of 30 hrs) and collectively form the coursework for a structured masters degree which also includes a research component. The advertised content is a guide, and the lecturers are encouraged, and indeed expected, to adapt daily to meet the current needs of the students.

Over the past ten years AIMS has achieved international recognition for this innovative and flexible approach. It has been the starting point for the remarkable success of our students and alumni and we all benefit from the support of many who have "witnessed the AIMS-magic and keep coming back for more."

This year we have decided to film selected courses and to make them available to a larger audience as an online facility. African universities may choose to use these courses to supplement and enhance their own postgraduate programmes. We believe this would be best achieved through engagement with AIMS. One way for this to happen, would be for AIMS to suggest or nominate a specialist tutor to spend time at the university, guiding students who follow the online programme. Where possible expert lecturers who have taught at AIMS may visit the university to give a short introduction to the course. We would welcome this interaction as well as the contribution our online courses will make to the growth of the mathematical sciences ecosystem in Africa.

Barry Green
Director & Professor of Mathematics
African Institute for Mathematical Sciences
January 2013

PROBABILITY & STATISTICS

2012

PROF DAVID SPIEGELHALTER

DAY 5

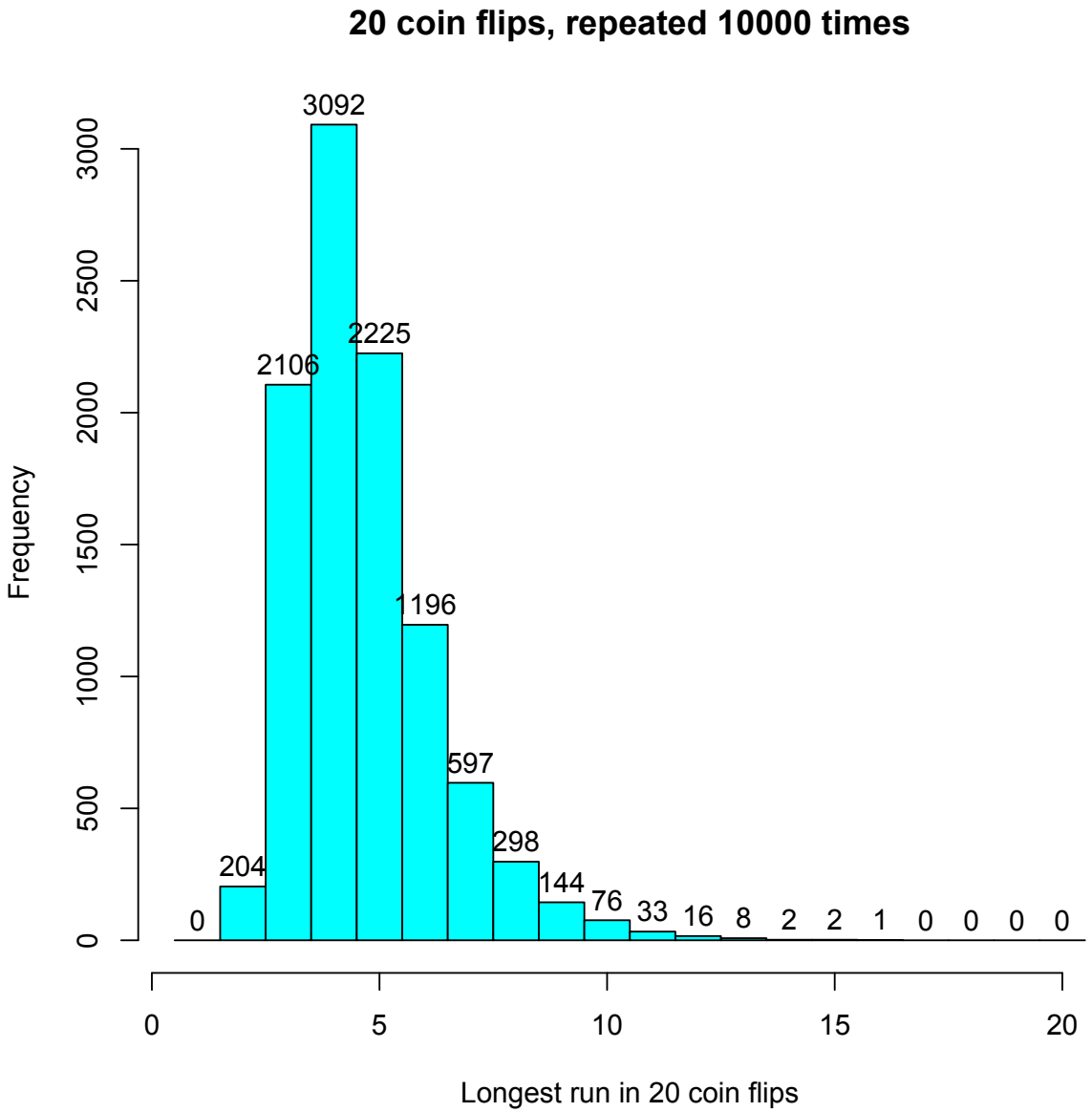
Distinguishing real from pretend coin flips

Suppose we observe a sequence of 20 coin flips - how can we tell if they are real or fake?

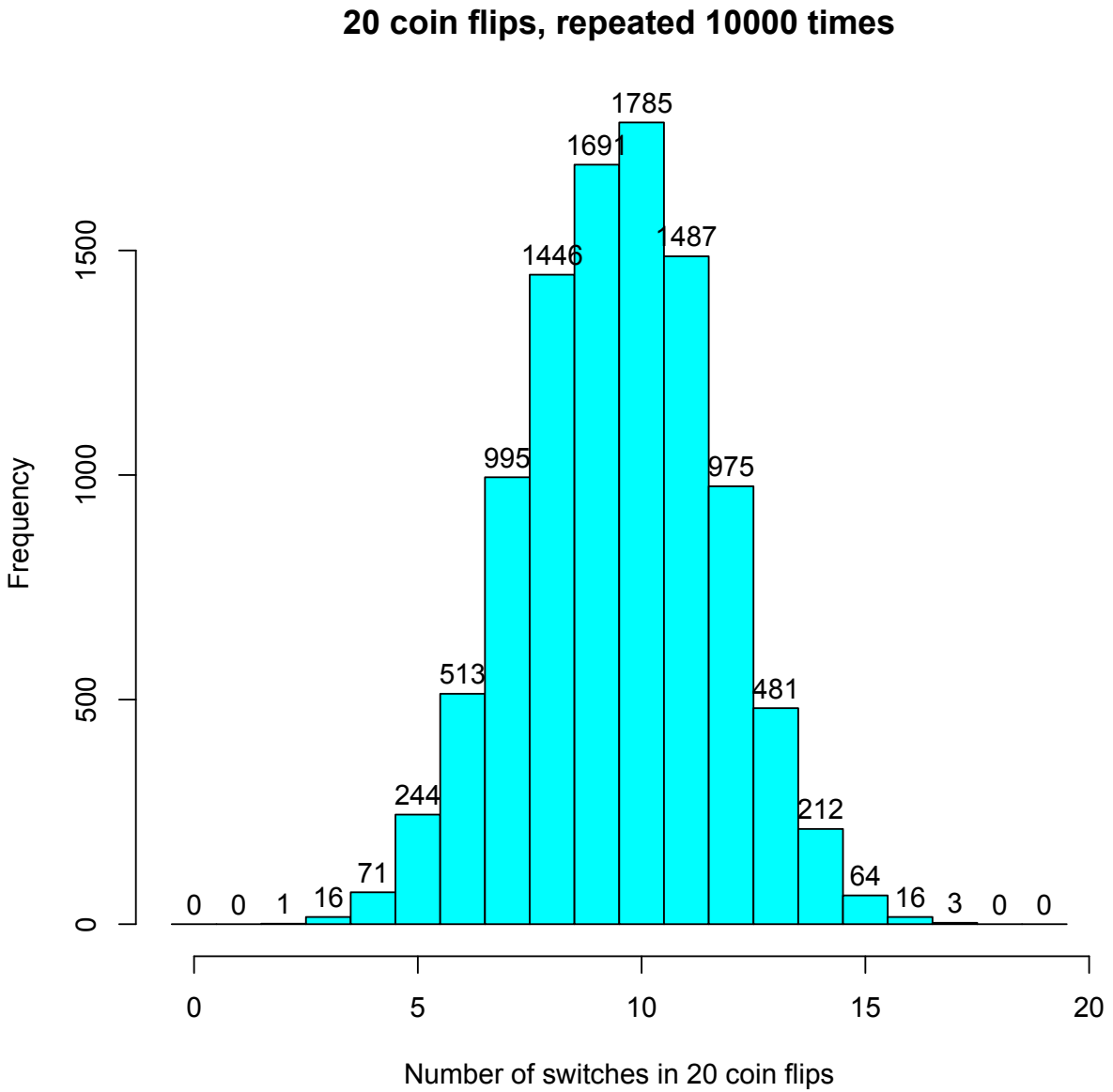
Two standard measures are

- Longest sequence of either heads or tails. We can get a rough idea by noting that a sequence of length 4 being all the same has probability 1/8, there are 16 sequences of length 4 (although clearly not independent as they overlap), so we might expect two runs of 4. Use simulation to get full distribution.
- The number of 'switches' between heads and tails. The waiting time till a switch is Geometric(0.5), so the gap between switches is 2. So we would expect 10 switches. Use simulation to get full distribution.

The longest run in 20 coin flips



The number of switches in 20 coin flips



Continuous random variables

Definition: a random variable X is continuous if

- $P(X = x) = 0$, for all x (i.e. there is probability 0 of taking on any x precisely, e.g. 3.7128464537383993..
- It has a distribution function $F(x) = P(X \leq x)$, such that $F(-\infty) = 0, F(\infty) = 1$
- $P(a < X < b) = F(b) - F(a)$ (whether $<$ or $\leq$ does not matter since continuous)

If $F(x_p) = p$, the x_p is known as the '100pth percentile'.

$x_{0.5}$ is the 50th percentile, also known as the *median*.

$x_{0.25}, x_{0.75}$ are known as the *quartiles*.

Probability density functions

If F is *absolutely continuous*, then

$$f(x) = F'(x) = \frac{d}{dx}F(x)$$

exists and is known as the probability density function, where

$$\int_{-\infty}^{\infty} f(t)dt = 1, \quad \int_{-\infty}^{x} f(t)dt = F(x).$$

Can think the probability that X is in a small interval $(x, x + dt)$ is $f(x)dt$.

Note: $f(x)$ can be > 1 !

Normal distributions

- Denoted $X \sim \text{Normal}(\mu, \sigma^2)$
- Standardised density: if $Z = (X - \mu)/\sigma$,

  $Z \sim \text{Normal}(0, 1)$: $f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} \quad -\infty < z < \infty$
- Standardised distribution function:

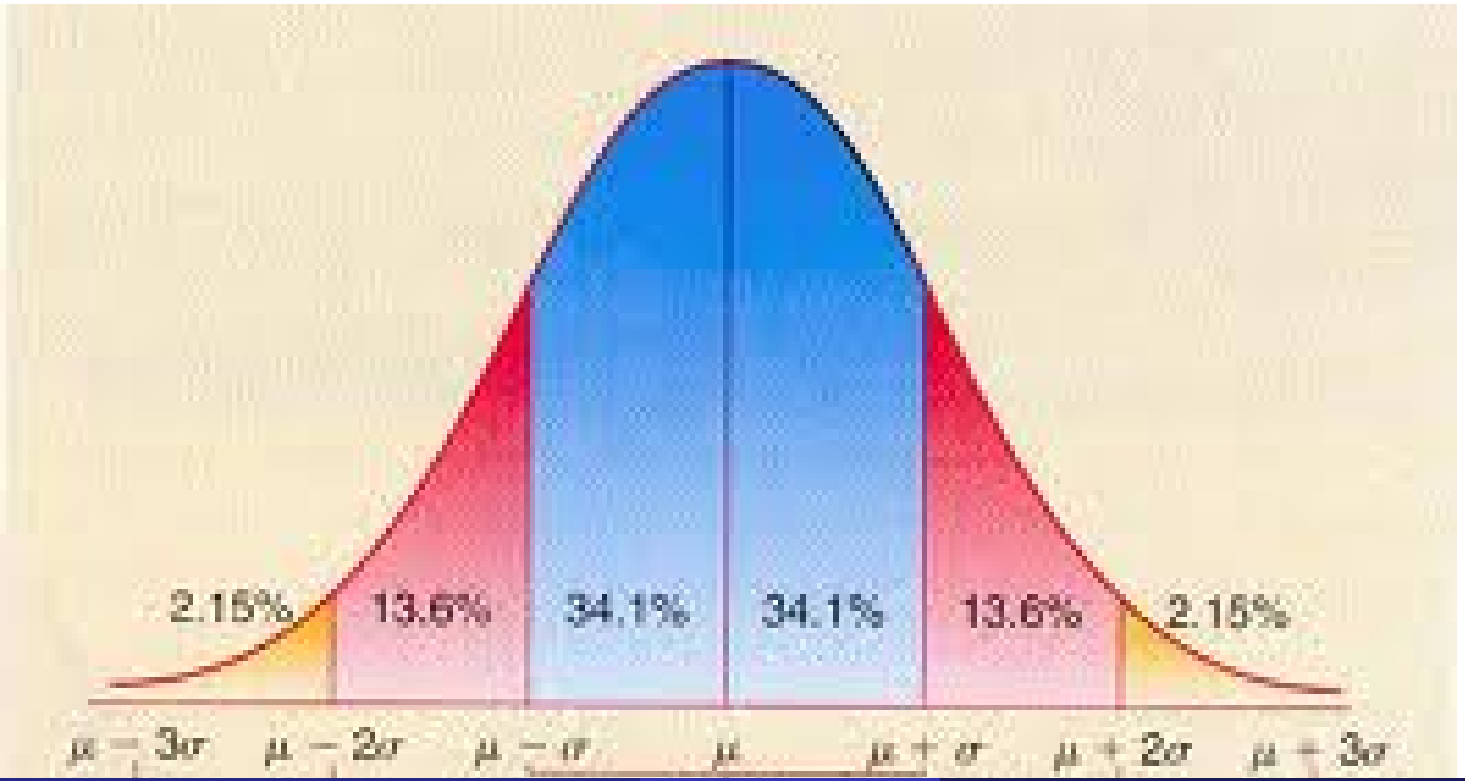
  $P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} dt = \Phi(z).$
- Density: $f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$
- Distribution function: $P(X < x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$
- Expectation (mean): μ
- Variance: σ^2
- Standard deviation $= \sigma$

Normal tail areas

To calculate areas, use

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) =$$

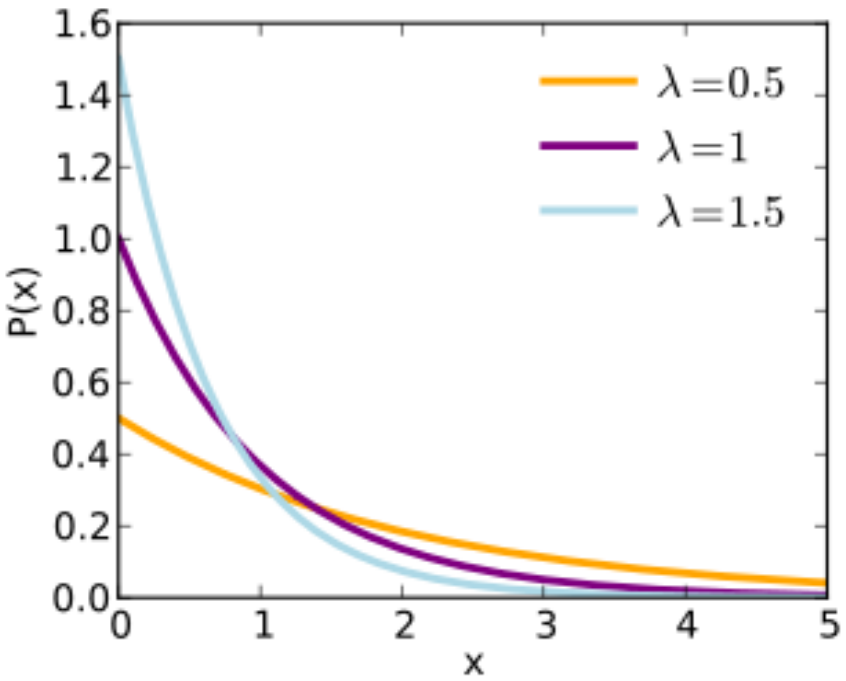
$$\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)$$



Exponential distribution

- Denoted $X \sim \text{Exp}(\lambda)$
- Standardised density: $Z \sim \text{Exp}(1): \ f(z) = e^{-z}, \ 0 < z < \infty$
- Standardised distribution function:
  $P(Z < z) = \int_{-\infty}^{z} e^{-t} dt = 1 - e^{-z}.$
- Density: $f(x|\lambda) = \lambda e^{-\lambda x}$
- Distribution function: $P(X < x) = 1 - e^{-\lambda x}.$
- Expectation (mean): $1/\lambda$
- Variance: $1/\lambda^2$
- Standard deviation $= 1/\lambda$

Exponential distribution



Used as a distribution for waiting times between random events

'Memoryless property': if the event has not happened by time t, then the distribution for the future waiting time is still the same exponential distribution

i.e. $(X - t)|(X > t)$ has an Exponential(λ) distribution

Uniform distribution

- Denoted $X \sim \text{Uniform}(a, b)$
- Density: $f(x|a, b) = \frac{1}{b-a}$; $a < x < b$
- Distribution function: $P(X < x) = 0$ if $x < a$; $(x - a)/(b - a)$ if $a < x < b$; 1 if $b < x$
- Expectation (mean): $(a + b)/2$
- Variance: $(b - a)^2/12$
- Standard deviation $= (b - a)/\sqrt{12}$

Inverse distribution function

If X has distribution function F, what is the distribution of $F(X)$?

$$P(F < f) = P(F(X) < f) = P(X < F^{-1}(f)) = F(F^{-1}(f)) = f$$

So the distribution function of any X has a uniform distribution!

This provides a means of sampling X: if F is available in closed form and can be inverted, then we

- Sample a random number u between 0 and 1
- Calculate $x = F^{-1}(u)$
- Then x is an observation from the distribution defined by F

Inverse distribution function

Example: how to sample from an exponential distribution with mean $1/\lambda$. $F(x|\lambda) = 1 - e^{-\lambda x}$, and so if we set $U = F(X) = 1 - e^{-\lambda X}$, this can be inverted to give

$$X = -\log(1 - U)/\lambda.$$

So we sample a uniform number U between 0 and 1, apply this transformation, and we have an observation from an exponential distribution.

Joint distributions for pairs of continuous random variables

The joint density for X, Y is given by

$$f(x,y) = \frac{\delta^2}{\delta x \delta y} F(x,y) = \frac{\delta^2}{\delta x \delta y} P(X \leq x, Y \leq y)$$

The conditional distribution for $X|Y$ is given by

$f(x|y) = f(x,y)/f(y)$ if $f(y)$ exists, 0 elsewhere.

We obtain the *marginal* distribution for X by integrating over the Y:

$$f(x) = \int f(x,y)dy = \int f(x|y)f(y)dy$$

'Extending the conversation'

They are said to be mutually independent if their joint density function is given by

$$f(x,y) = f(x)\ f(y).$$

Vectors of random variables

Let $\mathbf{X} = X_1, X_2, .., X_n$ be a vector of random variables.

The joint density for $\mathbf{X}$ is given by $f(x_1, .., x_n)$ We obtain the *marginal* distribution for a scalar X_1 by integrating (summing if discrete) over the remaining elements:

$$f(x_1) = \int f(x_1, .., x_n) dx_2 dx_3 .. dx_n.$$

They are said to be mutually independent if their joint density function is given by

$$f(x_1, .., x_n) = f(x_1)\ f(x_2), .., f(x_n).$$

Sums of random variables

Let $X_1, X_2, .., X_n$ be any random variables, not necessarily independent. Let $Y = \sum_{i=1}^{n} X_i$ be their sum. Then

$$\mathbb{E}[Y] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

So the expectation of a sum is always the sum of the expectations.

Variance of sums of independent random variables

Let $X_1, X_2, .., X_n$ be any *independent* random variables Let $Y = \sum_{i=1}^{n} X_i$ be their sum. Then

$$\mathbb{V}[Y] = \mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i]$$

So the variance of the sum is the sum of the variances *if independent*

Distribution of sums of independent discrete random variables

The Binomial(n, p) is the sum of n independent Bernoulli trials. Therefore it has mean np and variance $np(1-p)$.

The sum of n independent Poisson variables with means (expectations) $\mu_1, .., \mu_n$ is Poisson with mean $\mu_1 + .. + \mu_n$.

Can prove all these using *probability generating functions* (pgf)

Distribution of sums of independent continuous random variables

If $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$ are independent RVs, $i, 1, .., n$, then

$$Y = X_1 + X_2 + .. + X_n \sim \text{Normal}(\sum_i \mu_i, \sum_i \sigma_i^2)$$

So if all have the same distribution,

$$Y \sim \text{Normal}(n\mu, n\sigma^2)$$

And so the average $Y/n = \overline{X}$ has distribution

$$Y/n \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$$

Weak Law of Large Numbers

Averages of random variables tends to the mean

Let X_i be independent and identically distributed (iid) RVs, $i, 1, .., n$, with finite mean μ, and their mean be $\overline{X} = (X_1 + X_2 + .. + X_n)/n$. Then for any $\delta > 0$,

$$\lim_{n \to \infty} P\left(\left|\overline{X} - \mu\right| \geq \delta\right) = 0.$$

If X_i have a finite variance σ^2, then

$$P\left(\left|\overline{X} - \mu\right| \geq \delta\right) < \frac{\sigma^2}{n\delta^2}.$$

Central Limit Theorem

Distribution of averages of random variables tends to Normal around mean

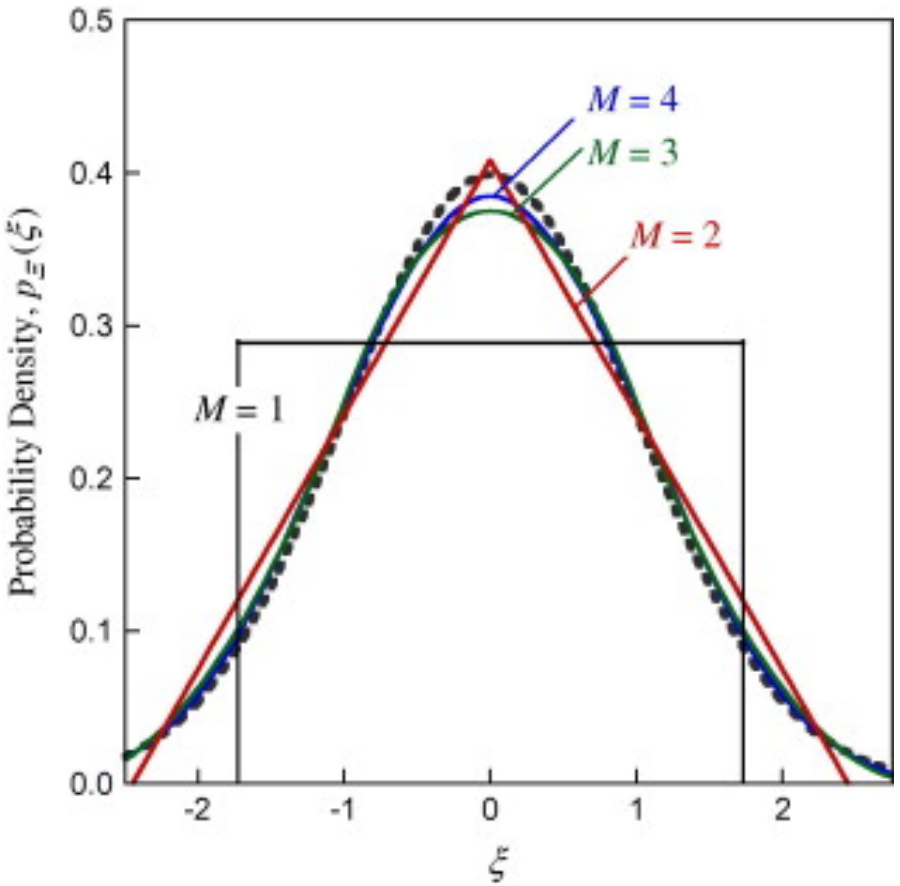
Let X_i be independent and identically distributed (iid) RVs, $i, 1, .., n$, with finite mean μ and finite variance σ^2, and their mean be $\overline{X} = (X_1 + X_2 + .. + X_n)/n$. Then

$$\lim_{n \to \infty} P\left(\frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x).$$

This is very powerful: it means that any average, based on a big enough sample, can be assumed to have a normal distribution.

Average of uniform variables

The distribution very quickly tends to a normal



Average of independent uniform variables

Sampling

We need to clearly distinguish between

- sampling X: if we graph the results, should follow the original distribution for X with standard deviation σ
- sampling $\overline{X}$: if we graph the results, should follow a Normal distribution with standard error $\sigma/\sqrt{n}$

e.g. Let X_i be a lottery number, n=6 on a ticket, and let $\overline{X}$ be the average lottery number drawn

Then many replicates of $\overline{X}$ should follow a normal distribution

Normal approximation to the Binomial

Let X_i be independent and identically distributed (iid) Bernoulli random variables, $i = 1, .., n$, each with mean p and variance $p(1-p)$.

Let $\overline{X} = (X_1 + X_2 + .. + X_n)/n$ be their average: the overall proportion of 'successes': then $\overline{X} = S_n/n$, where S_n has a Binomial(n, p) distribution.

By the Weak Law of Large Numbers, $\overline{X}$ tends to p.

By the Central Limit Theorem,

$$\overline{X} \sim \text{Normal}(p, \frac{p(1-p)}{n})$$

or equivalently

$$S_n \sim \text{Normal}(np, np(1-p))$$

So the Binomial can be approximated by the Normal - can also show this directly using Stirling's approximation.

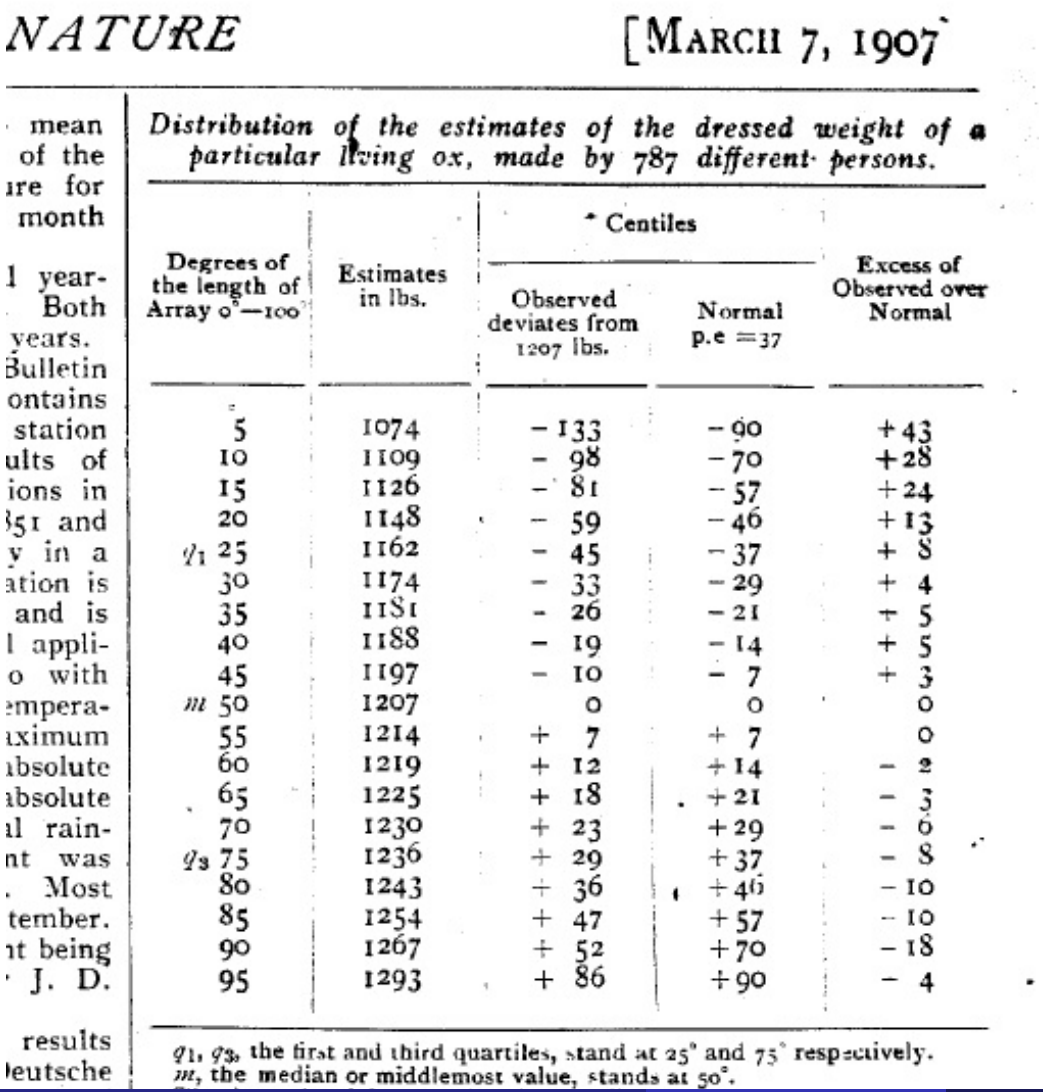
Example

A parliament of size 100 is supposed to be based on equal opportunities, but there are 65 men in the parliament, If 'equal opportunity' is taken to mean an equal chance that a seat will be held by a man or a woman, what is the chance of getting such an extreme result?

Let X be the number of males. Under equal opportunities,
$X \sim \text{Binomial}(100, 0.5)$, with mean 50 and variance 100
$p(1-p) = 100 \times 0.5 \times 0.5 = 25$, i.e. standard deviation 5. So
$P(X \geq 65) \approx P(Z > \frac{65-50}{\sqrt{25}}) = P(Z > 3) = 1 - \Phi(3) = 0.001$. So there is approximately only 1 in 1000 chance of getting such an imbalance by chance alone.

'Wisdom of the Crowds'

In 1907 Francis Galton obtained 787 guesses of the weight of a butchered ox. True weight: 1198 lb



NATURE [MARCH 7, 1907

mean
of the
ire for
month

l year-
Both
years.
Bulletin
ontains
station
ults of
ions in
351 and
y in a
ation is
and is
l appli-
o with
empera-
aximum
absolute
absolute
al rain-
nt was
. Most
tember.
nt being
J. D.

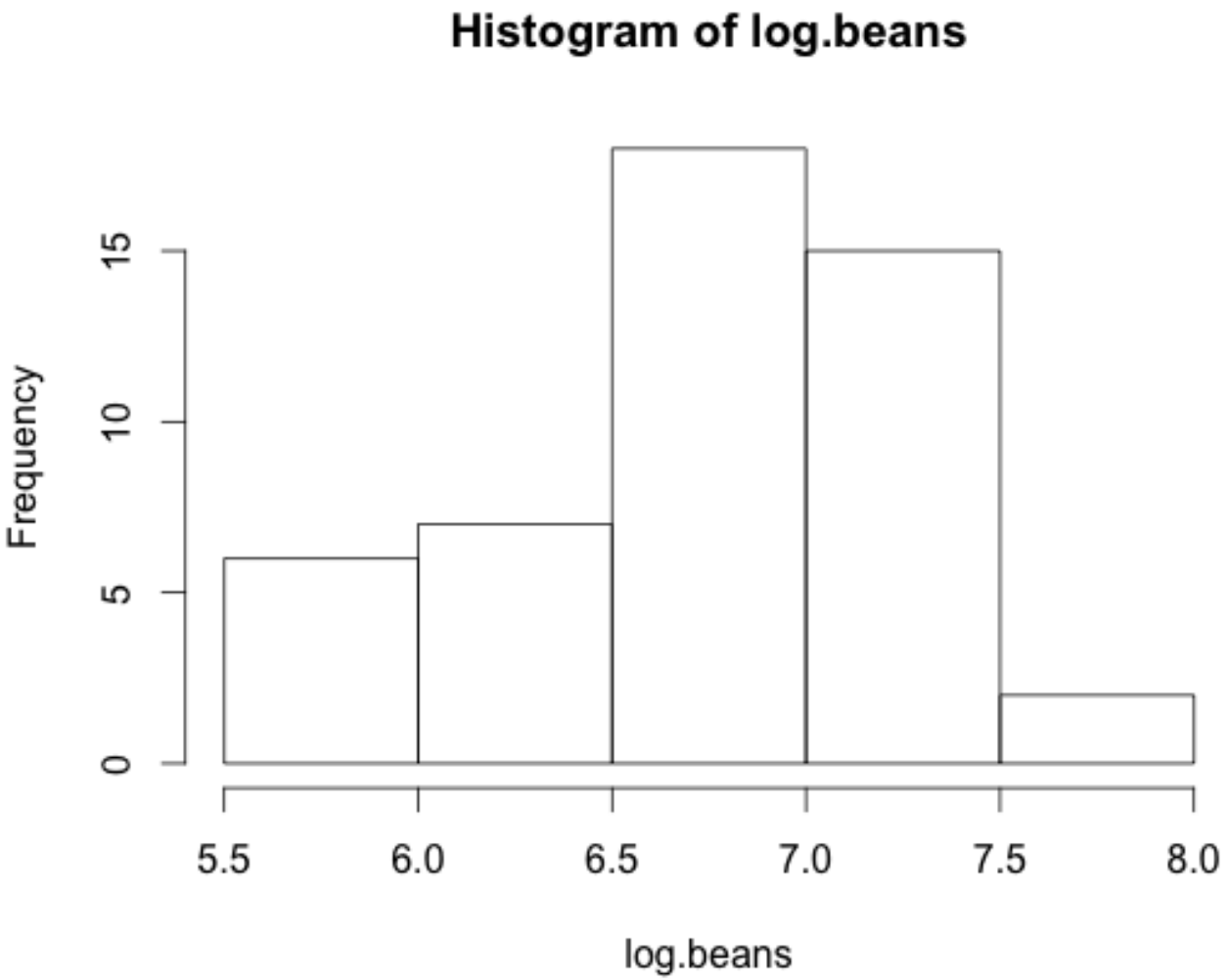
results
leutsche

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

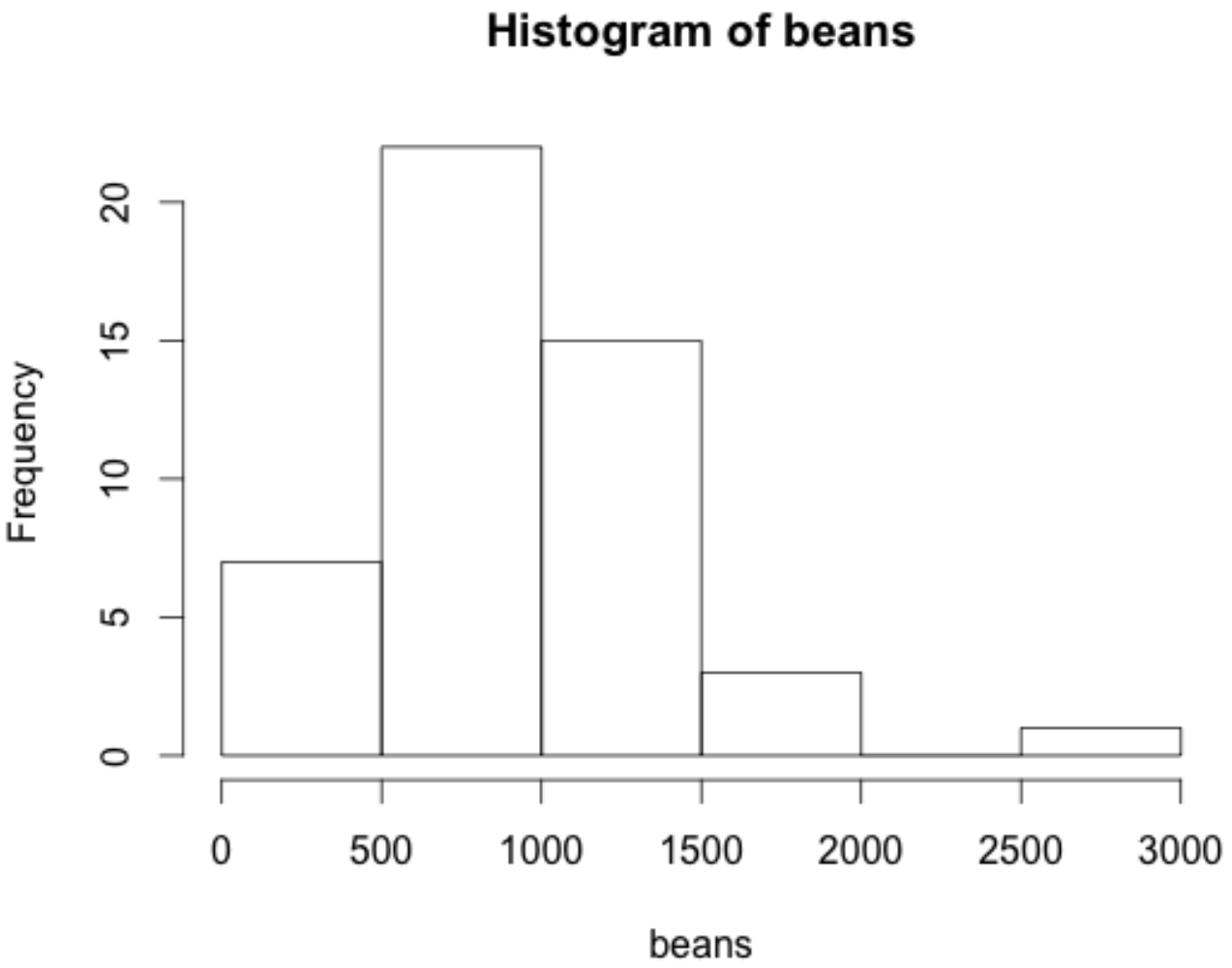
q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.

log(Bean counts)



More symmetric. A variable whose logarithm is Normal, is known as having a *log-normal distribution*

Bean counts



Mean 977, median 925. The truth???

Assignment 1

Please hand in solutions to these questions, preferably in Latex

1. The genetic code specifies an amino acid by a sequence of three nucleotides. Each nucleotide can be one of four kinds: T, A, C or G, with repetitions permitted. How many amino acids can be coded in this manner? How would the answer change if repetitions were not allowed?

2. Experience shows that 10% of people who make reservations for a plane trip do not show up. An airline takes 100 reservations - what is the distribution for the number of people who will show up? What is its mean and variance?

   The plane has 90 seats. What is the probability that the plane will be overbooked and someone will not be able to travel? [This can be calculated exactly (using a suitable program) or using a Normal approximation]

3. The number of goals scored by a football team in each match has a Poisson distribution with mean 1. After 20 games, what is the mean and variance for the total number of goals scored?

4. if X has an Exponential distribution Exp(λ), show that the variance of X is $1/\lambda^2$.

   [You may want to use the definition of the Gamma function: $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt = (z-1)!$]

The following R questions do not need to be handed in.

1. Translate into R the previous Binomial dice throwing program (1000 simulations of throwing 6 dice and counting the number of 6's that appear). Use the `barplot()` command to make bar charts of the true and simulated distributions.

   Websites such as `http://www.harding.edu/fmccown/r/` give the simplest commands, and show how to make the graphs more pretty.

2. Read in the class bean-counting data from the website. Then you can get the genders and counts from

   Create a histogram for the counts for the combined groups, and separately for men and women. Calculate the sample mean and variance of the counts within each gender.

   Do you think there is a difference between the genders?