STATS 3N03/3J04

2004-10-06
10-1

RANDOM VARIABLE X

SAMPLE SPACE $\mathcal{X}$

PARTICULAR VALUE $x \in \mathcal{X}$

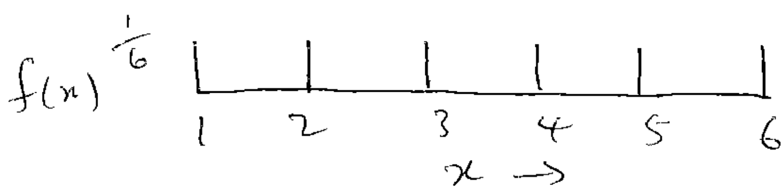
PROBABILITY MASS FUNCTION
(PROBABILITY DENSITY FUNCTION)

$$f(x) = P(X = x)$$

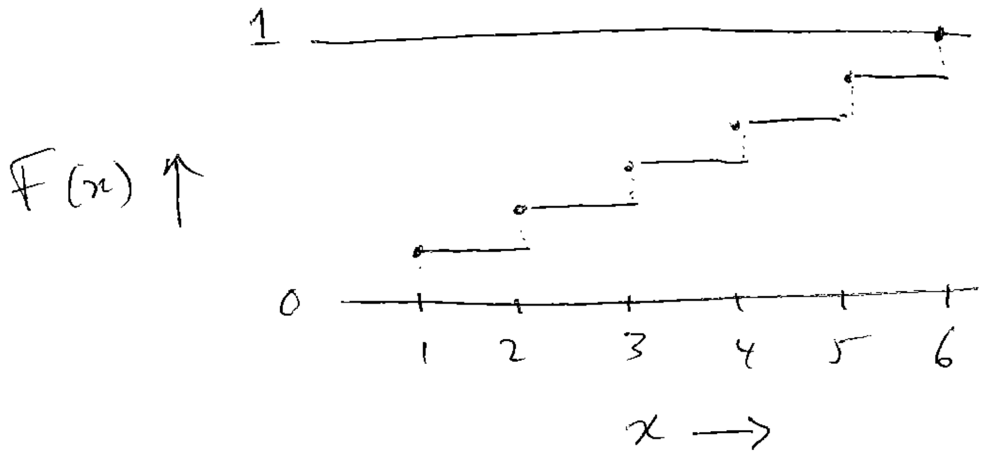
CUMULATIVE DISTRIBUTION FUNCTION

$$F(x) = P(X \leq x)$$

EX



x	1	2	3	4	5	6
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{6}{6}$



MEAN / AVERAGE / EXPECTED VALUE

$$\mu = E(X) = \sum_{x \in \mathcal{X}} x f(x)$$

$$= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= \frac{21}{6} = 3.5$$

- CENTRE OF MASS OF PDF
- WEIGHTED AVERAGE OF ALL VALUES X MIGHT TAKE
- IN THE LONG RUN, THE SAMPLE MEAN $\bar{x}$ WILL CONVERGE TO μ, AS n GETS LARGE.

NOTE: CONVERGENCE IS IN A PROBABILISTIC SENSE, FOR ANY GIVE INTERVAL ABOUT μ, THE CHANCE THAT $\overline{x}$ WILL LIE OUTSIDE THAT INTERVAL GOES TO 0 AS n GOES TO ∞.

VARIANCE / MEAN SQUARED DEVIATION

$$\sigma^2 = \text{Var}(X) = E\left[(X-\mu)^2\right] = \sum_{x \in \mathcal{X}} (x-\mu)^2 f(x)$$

$$= (1-3.5)^2 \tfrac{1}{6} + (2-3.5)^2 \tfrac{1}{6} + (3-3.5)^2 \tfrac{1}{6} + (4-3.5)^2 \tfrac{1}{6} + (5-3.5)^2 \tfrac{1}{6} + (6-3.5)^2 \tfrac{1}{6}$$

$$= (-2.5)^2 \tfrac{1}{6} + (-1.5)^2 \tfrac{1}{6} + (-0.5)^2 \tfrac{1}{6} + (.5)^2 \tfrac{1}{6} + (1.5)^2 \tfrac{1}{6} + (2.5)^2 \tfrac{1}{6}$$

$$= \frac{17.5}{6} = 2.916\dot{6}$$

INTERPRETATION: MOMENT OF INERTIA ABOUT THE CENTRE OF MASS $\Rightarrow$ MEASURE OF DISPERSION IN UNITS OF X^2

STANDARD DEVIATION

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)}$$

$$= \sqrt{2.91\dot{6}} = 1.7078$$

- IN THE LONG RUN THE SAMPLE VARIANCE s^2 WILL CONVERGE TO σ^2, AS n GETS LARGE.

THEORETICAL: $f(x)$ $F(x)$ μ σ^2

OBSERVED: $\hat{f}$ (HISTOGRAM) $F_n(x)$ $\bar{x}$ s^2

THEORETICAL IS A <u>MODEL</u> FOR OBSERVED <u>DATA</u>.

OBSERVED DATA ESTIMATES THE THEORETICAL MODEL.

EX DATA FROM 20 ROLLS OF A FAIR 6-SIDED DIE.

5, 2, 4, 4, 6, 6, 3, 3, 3, 4, 6, 4, 2, 4, 3, 3, 6, 4, 5, 1

$$\bar{x} = \frac{1}{20}(5+2+\cdots+1) = \frac{78}{20} = 3.9$$

$$s_x^2 = \frac{1}{19}\left((5-3.9)^2 + (2-3.9)^2 + \cdots + (1-3.9)^2\right)$$

$$= 2.094737$$

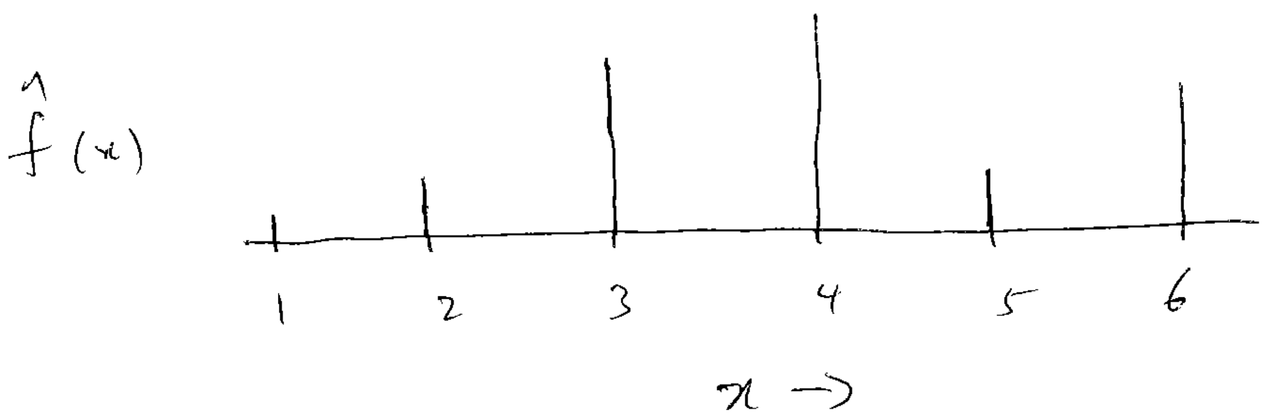
EASIER: $\sum x_i = 78$ $\sum x_i^2 = 344$

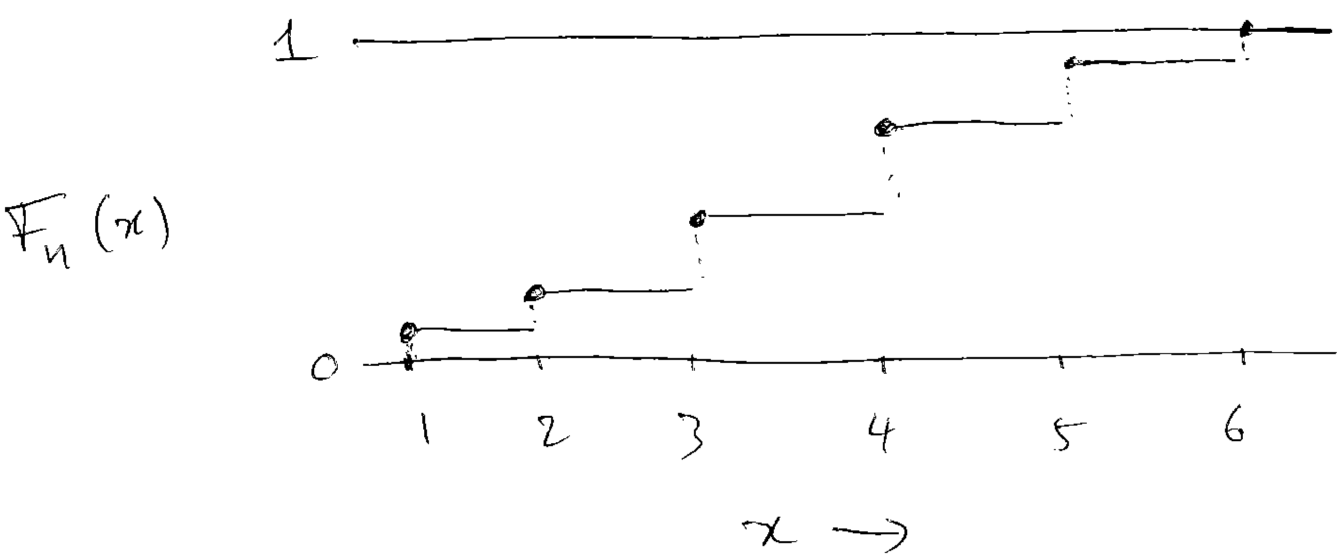
$$s_x^2 = \frac{1}{19}\left(344 - \frac{78^2}{20}\right) = \frac{39.8}{19}$$

$$= 2.094737$$

GROUPED DATA

x	1	2	3	4	5	6
freq	1	2	5	6	2	4
$\hat{f}(x)$	.05	.10	.25	.30	.10	.20
$F_n(x)$	.05	.15	.40	.70	.80	1.00





SAMPLE MEAN FROM GROUPED DATA:

$$\bar{x} = \sum_{x \in \mathcal{X}} x \hat{f}(x)$$

$$= 1\frac{1}{20} + 2\frac{2}{20} + 3\frac{5}{20} + 4\frac{6}{20} + 5\frac{2}{20} + 6\frac{4}{20}$$

$$= \frac{78}{20} = 3.9 \qquad (\mu = 3.5)$$

SAMPLE VARIANCE FROM GROUPED DATA:

$$s_x^2 = \frac{n}{n-1}\sum (x - \bar{x})^2 \hat{f}(x)$$

ADJUSTMENT TO REMOVE BIAS

$$= \frac{20}{19}\left\{(1-3.5)^2\frac{1}{20} + (2-3.5)^2\frac{2}{20} + \cdots + (6-3.5)^2\frac{4}{20}\right\}$$

$$= \cdots = 2.094737$$

$$(\sigma^2 = 2.91\dot{6})$$

GROUPED DATA IN R

x ← 1:6

freq ← c(1, 2, 5, 6, 2, 4)

$\bar{x}$ mean(rep(x, freq))

s_x^2 var (rep(x, freq))

s_x sqrt(var(rep(x, freq)))

i.e. USE rep( ) TO UNGROUP.

NOTE:

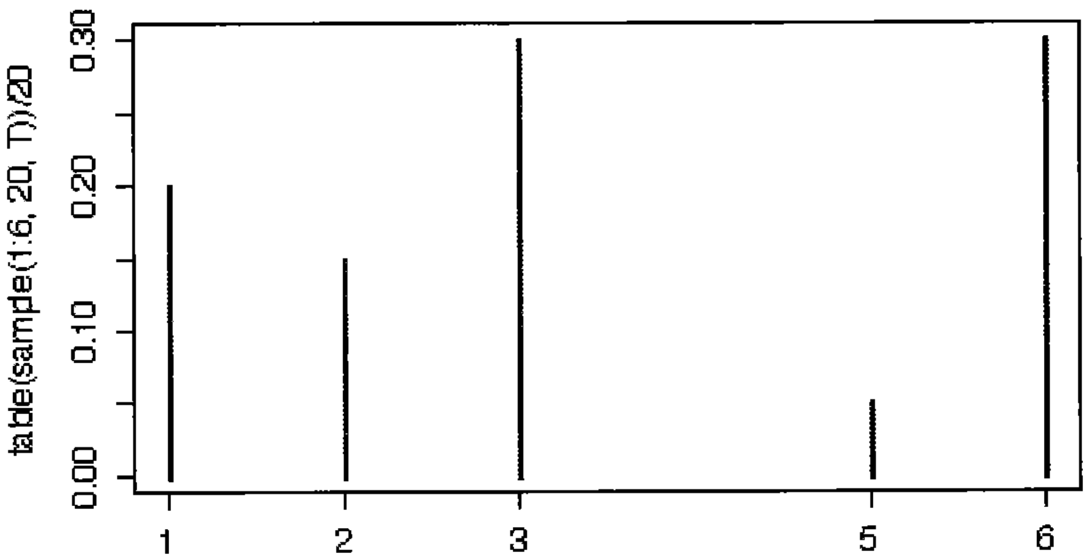
n = 20 SEEMS OK TO ESTIMATE μ AND σ^2

BUT ESTIMATE OF PDF ISN'T GOOD ENOUGH.

HOW LARGE SHOULD n BE ??

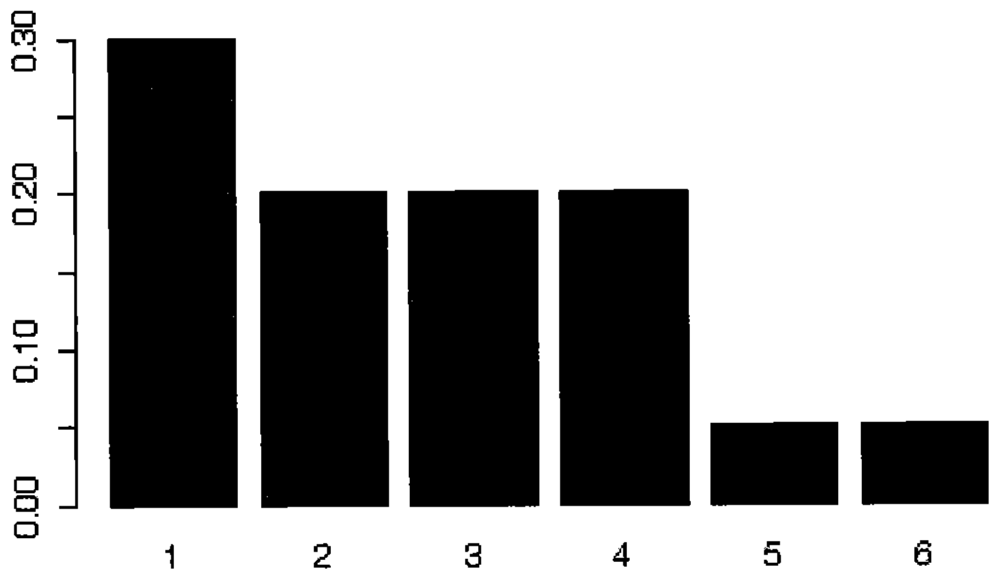
10-9

```
> sample(x=1:6, size=20, replace=T)
 [1] 2 5 5 4 6 6 1 5 5 1 1 1 2 6 1 5 5 6 2 1

> sample(x=1:6, size=20, replace=T)
 [1] 5 5 6 6 6 2 1 4 1 5 3 5 4 3 3 3 2 6 4 4

> table(sample(x=1:6, size=20, replace=T))

1 2 3 4 5 6
3 2 4 2 7 2


> plot(table(sample(1:6,20,T))/20, col="blue")
```



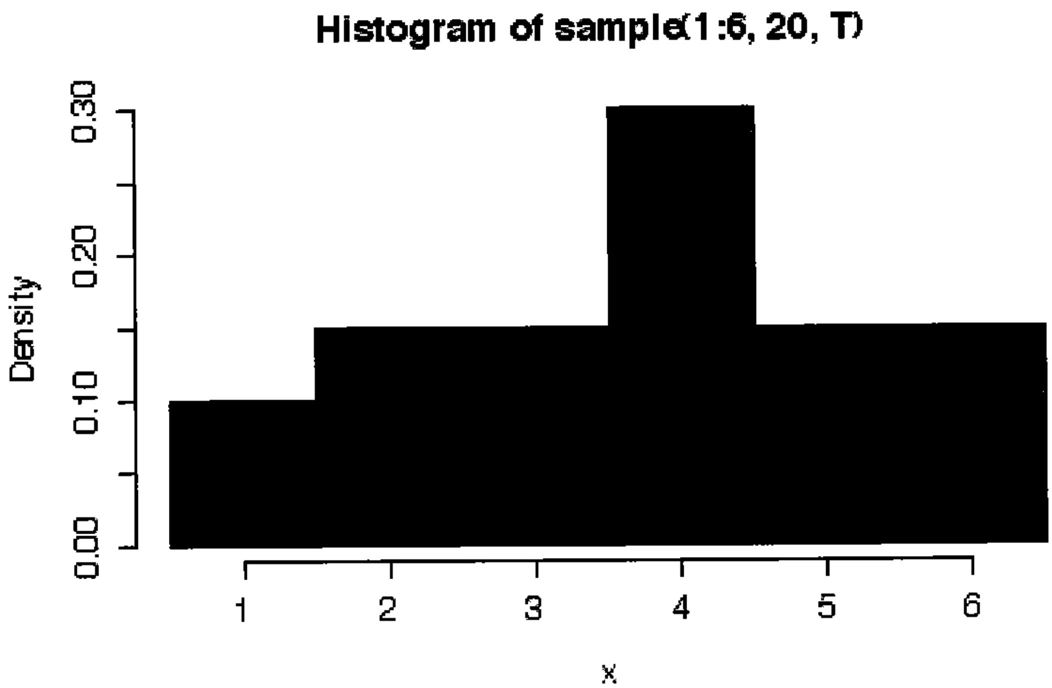
10-10

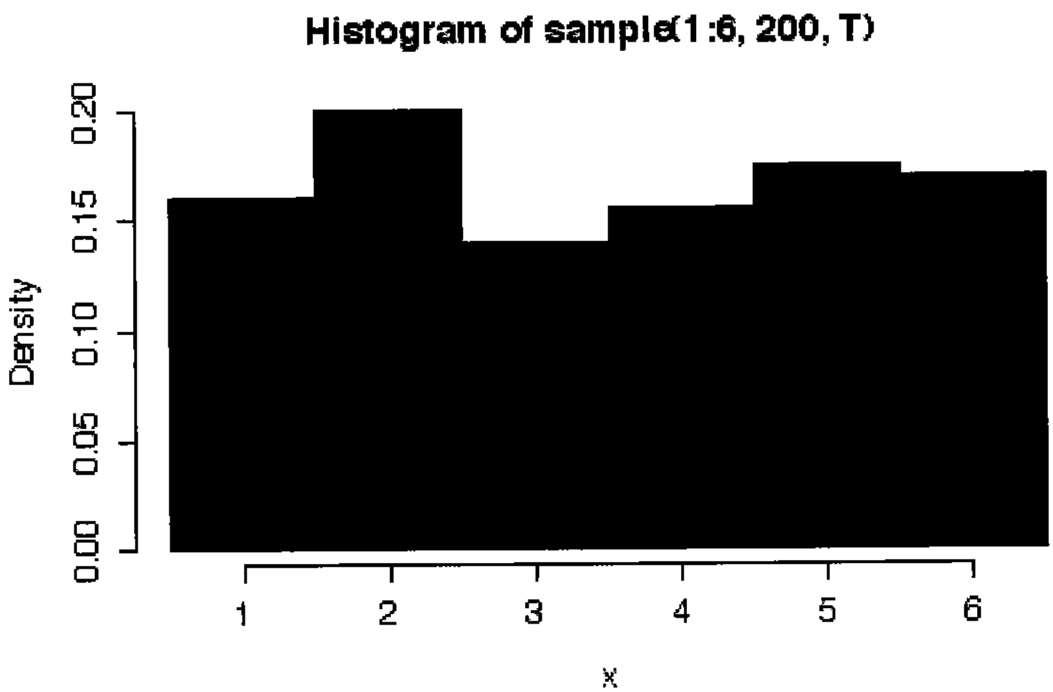
```
> barplot(table(sample(1:6,20,T))/20, col="blue")
```



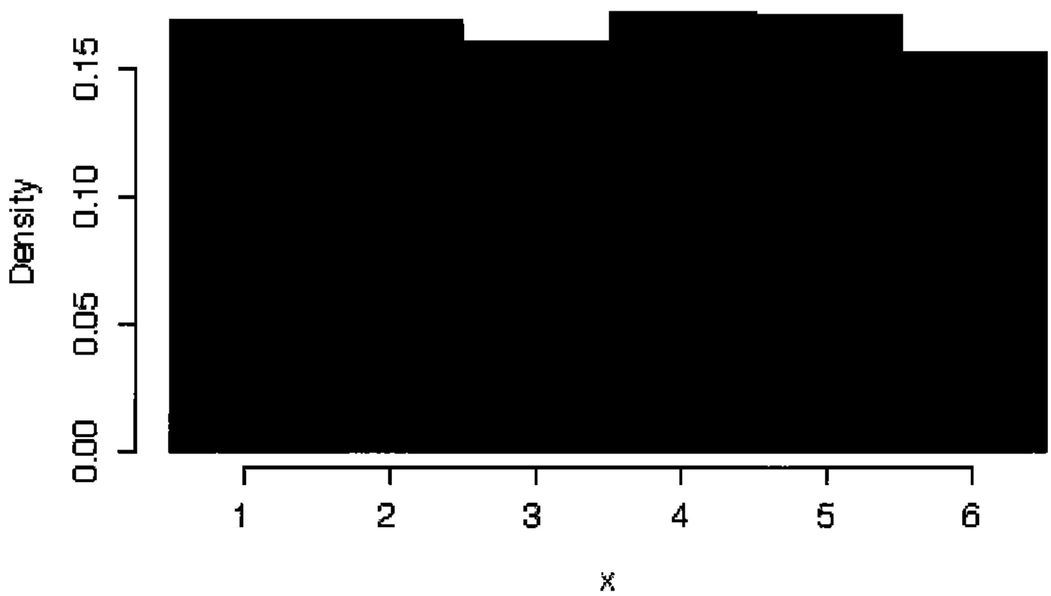
10-11

```
> hist(sample(1:6,20,T), prob=T, breaks=seq(.5,6.5,by=1),
col="blue", xlab="x")
> hist(sample(1:6,200,T), prob=T, breaks=seq(.5,6.5,by=1),
col="blue", xlab="x")
> hist(sample(1:6,2000,T), prob=T, breaks=seq(.5,6.5,by=1),
col="blue", xlab="x")
> hist(sample(1:6,200000,T), prob=T, breaks=seq(.5,6.5,by=1),
col="blue", xlab="x")
```

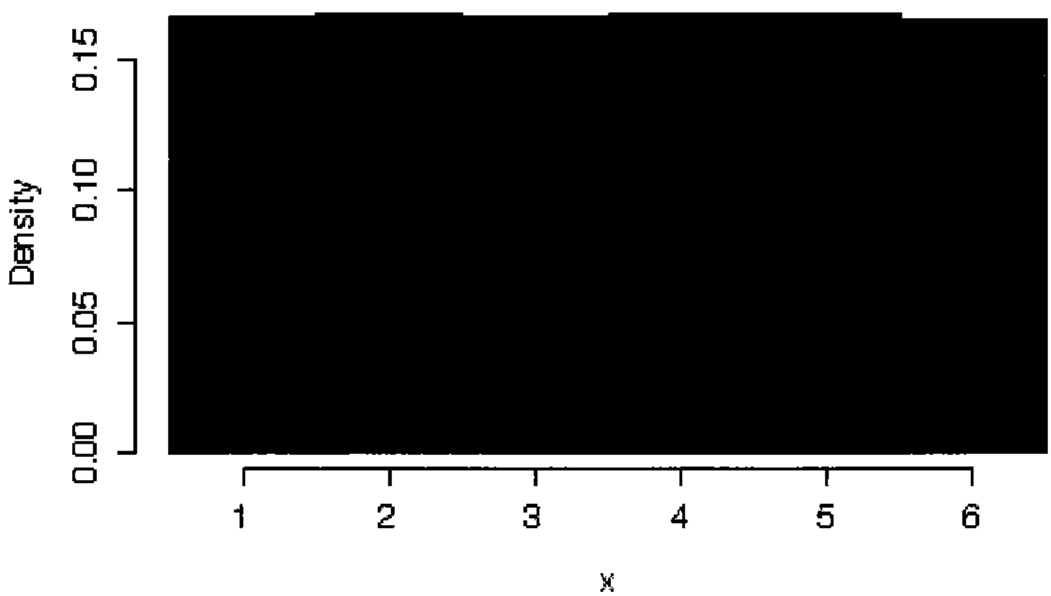




Histogram of sample(1:6, 2000, T)



Histogram of sample(1:6, 2e+05, T)



BINOMIAL DISTRIBUTION

EX RANDOM VARIABLE

X = NUMBER OF 1s IN 4 INDEPENDENT ROLLS OF A FAIR DIE

$\mathcal{X} = \{0, 1, 2, 3, 4\}$

FIND $f(x)$

OUTCOME	PROB.	X	$f(x)$
FFFF	$\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4$	0	$\binom{4}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4$
TFFF, FTFF, FFTF, FFFT	$\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$	1	$\binom{4}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3$
TTFF, TFTF, TFFT, FTTF, FTFT, FFTT	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2$	2	$\binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2$
TTTF, TTFT, TFTT, FTTT	$\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$	3	$\binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1$
TTTT	$\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^0$	4	$\binom{4}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^0$

GENERAL FORMULA:

n INDEPENDENT TRIALS OF A CHANCE SET-UP, EACH TRIAL CAN RESULT IN "SUCCESS" OR "FAILURE", CHANCE OF SUCCESS ON ANY GIVEN TRIAL IS p.

LET X = TOTAL NUMBER OF SUCCESSES

$$f(x) = P(X=x) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

$$x = 0, 1, \cdots, n$$

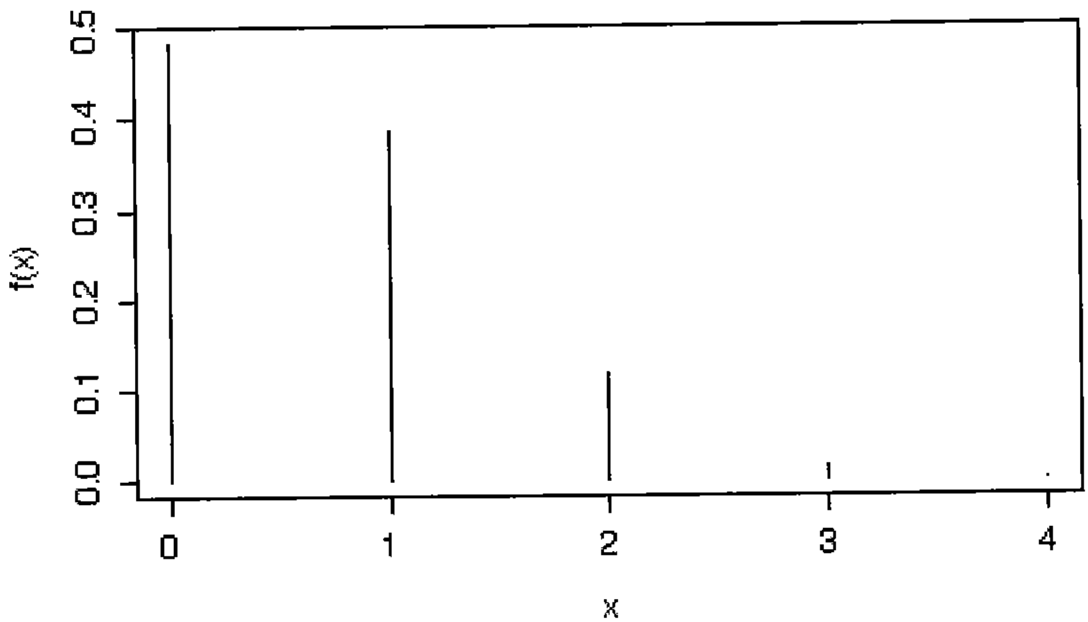
NOTE: IF $q = 1-p$

$$(p+q)^{n} = \sum_{x=0}^{n} \binom{n}{x} p^{x} q^{n-x}$$

"BINOMIAL EXPANSION"

10-15

```
> dbinom(x=0:4, size=4, prob=1/6)
[1] 0.482253086 0.385802469 0.115740741 0.015432099 0.000771605

> plot(0:4, dbinom(0:4, 4, 1/6), type="h", xlab="x",
ylab="f(x)")
```




```
> plot(0:100, dbinom(0:100, 100, 1/2), type="h", xlab="x",
ylab="f(x)")
```