

## Binomial Distribution

### Overview

- Recall that a r.v. is discrete if there are either a finite number of possible values (e.g.  $y_1, \dots, y_n$ ) or at most there is one for every integer (e.g.  $y_1, y_2, \dots$ ).
- Also recall that the probability distribution of a discrete r.v. is described by the probability of each possible value of the r.v. (i.e.  $p(y) = P(Y = y)$  for all possible  $y$ ).
- Some particular probability distributions occur often because they are useful descriptions of certain chance phenomenon under study.
- For example, binomial distribution for discrete r.v. is very useful in practice.

56

## Binomial Distribution

### Example: coin tossed three times

Toss a coin *independently* three times and record the number of times that the coin landed on heads. Let  $Y = \#$  of heads. Then the sample space is:

1st toss	2nd toss	3rd toss	Y
H	H	H	3
H	H	T	2
H	T	H	2
H	T	T	1
T	H	H	2
T	H	T	1
T	T	H	1
T	T	T	0

Suppose  $p =$  probability of heads and  $q =$  probability of tails. Then  $q = 1 - p$  and

$$P(Y = 3) = p \times p \times p = p^3$$

$$P(Y = 2) = p \times p \times q + p \times q \times p + q \times p \times p = 3p^2q$$

$$P(Y = 1) = 3pq^2$$

$$P(Y = 0) = q^3$$

57

## Binomial Distribution

### Definition

A *binomial distribution* must satisfy three (3) conditions:

1. There are  $n$  trials, each of which can result in one of the two outcomes, either “success” or “failure”. (i.e. the trials are dichotomous or *Bernoulli*).
2. The probability of a success is constant  $p$  for all trials.
3. The trials are independent.

Define r.v.  $Y = \#$  of successes in  $n$  trials. Then  $Y$  is said to have a *binomial distribution with parameters  $n$  and  $p$*  and is written as:

$$Y \sim B(n, p)$$

58

## Binomial Distribution

### Formula

Let  $Y \sim B(n, p)$ . The formula for the binomial distribution is:

$$P(Y = r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where

- $r = 0, 1, 2, \dots, n$  are the possible numbers of successes.
- $q = 1 - p$  is the probability of failure.
- $p^r = p \times p \times \dots \times p$  ( $r$  times).
- $q^{n-r} = q \times q \times \dots \times q$  ( $n - r$  times).
- $\binom{n}{r} = \frac{n!}{r!(n-r)!}$  is read as “ $n$  choose  $r$ ”.
- A quick review of factorial operation

$$k! = k \times (k - 1) \times \dots \times 1$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

$$0! = 1.$$

59

## Binomial Distribution

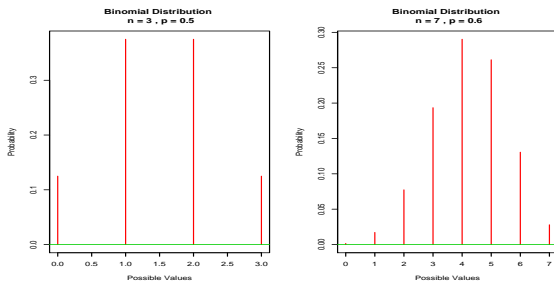
### Examples

- Suppose  $Y \sim B(3, \frac{1}{2})$ .

$$P(Y = 1) = \frac{3!}{1!2!} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

- Suppose  $Y \sim B(7, 0.6)$ .

$$\begin{aligned} P(Y = 3) &= \frac{7!}{3!4!} (0.6)^3 (0.4)^4 \\ &= \frac{7 \times 6 \times 5}{3!} (0.6)^3 (0.4)^4 = 0.194 \end{aligned}$$



60

## Binomial Distribution

### One more example

- Toss a fair die three times. What is the probability that there is at least once a 6?  
Let  $Y \sim B(3, \frac{1}{6})$ . Then

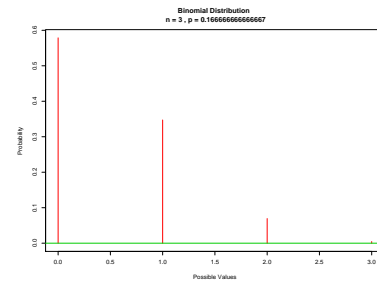
$$P(\text{at least one 6}) = 1 - P(\text{no 6}) = 1 - P(Y = 0).$$

Since

$$P(Y = 0) = \frac{3!}{0!3!} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3 = 0.579$$

we have

$$P(\text{at least one 6}) = 1 - 0.579 = 0.421.$$



61

## Binomial Distribution

### Properties

- $Y \sim B(n, p)$  is a discrete r.v.
- The expectation of  $Y$  is

$$\mu_Y = E(Y) = np$$

- The variance of  $Y$  is

$$\sigma_Y^2 = Var(Y) = npq$$

- For example, suppose  $Y \sim B(10, \frac{1}{4})$ ,

$$\begin{aligned} E(Y) &= 10 \times \frac{1}{4} = \frac{10}{4} = 2.5, \\ Var(Y) &= 10 \times \frac{1}{4} \times \frac{3}{4} = \frac{30}{16}. \end{aligned}$$

- Another example, suppose  $Y \sim B(100, \frac{1}{4})$ ,

$$\begin{aligned} E(Y) &= 100 \times \frac{1}{4} = \frac{100}{4} = 25, \\ Var(Y) &= 100 \times \frac{1}{4} \times \frac{3}{4} = \frac{300}{16}. \end{aligned}$$

62

## Binomial Distribution

### Proportions

In some situations, one is interested in the *proportion* of successes among  $n$  trials. Define

$$W = \frac{Y}{n}$$

where  $Y \sim B(n, p)$ . Then

$$E(W) = p, \quad Var(W) = \frac{pq}{n}$$

using properties of binomial distribution (i.e. expectation and variance of  $Y$ ).

63

## Binomial Distribution

### Remarks

- Binomial distribution is a very important probability model and is often very useful.
- For example, cure rate of a new drug, proportion of lakes that are suitable habitats for fish, survival rate of seedlings, etc.
- However, binomial distribution is not suitable for all situations. Are the trials dichotomous? Is the success probability constant? Are the trials independent?
- Consider the fair die game, # of rats cured within the same litter, etc.
- Always think about the three assumptions before using a binomial distribution.

64

## Normal Distribution

### Overview

- Recall that a r.v. is discrete if there are either a finite number of possible values (e.g.  $y_1, \dots, y_n$ ) or at most there is one for every integer (e.g.  $y_1, y_2, \dots$ ).
- In contrast a r.v. is continuous if the possible values are over an entire interval of numbers.
- The probability distribution of a continuous r.v. is described by a probability density curve such that area under the curve corresponds to probability.
- The most frequently used type of continuous distribution is normal distribution, which is also known as Gaussian distribution.
- Normal distribution is a very important distribution for modeling continuous data, because data that are influenced by many small and unrelated random effects are approximately normally distributed (e.g. a student's height).

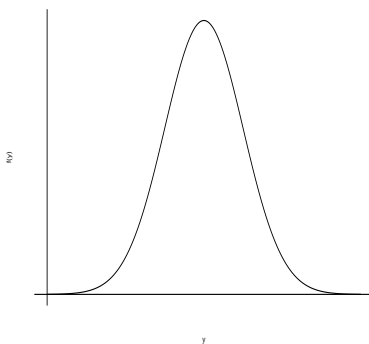
65

## Normal Distribution

### Definition

A *normal distribution* is defined by a bell-shaped curve with a probability density function

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$



66

## Normal Distribution

### Definition and properties

- A r.v.  $Y$  is said to have a *normal distribution with parameters  $\mu$  and  $\sigma^2$*  and is written as:

$$Y \sim N(\mu, \sigma^2)$$

- Total area under the distribution curve is 1.
- The distribution curve is symmetric about the center  $\mu$ . Values around  $\mu$  are more probable.
- The area in the range of  $\mu \pm \sigma$  is about 2/3 of the total area (0.6826).
- The distribution curve has a well-defined bell shape. The parameters  $\mu$  and  $\sigma^2$  provide complete information about the distribution curve.
- In fact,

$$E(Y) = \mu, \quad Var(Y) = \sigma^2.$$

67

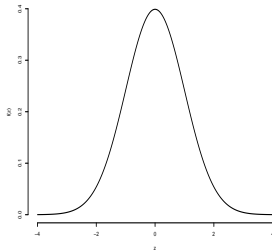
## Standard Normal Distribution

### Definition and properties

- A r.v.  $Z$  has a *standardized normal distribution* if  $Z \sim N(\mu, \sigma^2)$  with expectation  $\mu = 0$  and variance  $\sigma^2 = 1$  and is written as:

$$Z \sim N(0, 1).$$

- The distribution curve is symmetric about the center  $\mu = 0$ .
- The area between  $-1$  and  $1$  (i.e.  $\mu \pm \sigma$ ) is 0.6826 (about  $2/3$ ).
- The area between  $-2$  and  $2$  (i.e.  $\mu \pm 2\sigma$ ) is 0.9545.



68

## Standard Normal Distribution

### More properties

Suppose  $Z$  is a standard normal r.v. (i.e.  $Z \sim N(0, 1)$ ).

- $P(Z \leq 0) = 0.5$ .
- $P(-\infty < Z < \infty) = 1$ .
- $P(Z = 1) = 0$  and in general  $P(Z = c) = 0$  for any  $c$ .
- Hence  $P(Z < 0) = P(Z \leq 0) = 0.5$ .
- Remark: Suppose  $Y \sim B(3, 0.5)$  then

$$P(Y \geq 0) \neq P(Y > 0)!$$

- How to find  $P(Z \geq z)$  for any given value  $z$ ?

69

## Standard Normal Distribution

### Finding probabilities

- Table A on page 408 of the course notes (bluebook) gives

$$P(Z \geq z)$$

for any given value  $z \geq 0$  (also known as *the upper tail probability*).

- The  $z$  values are on the outside of Table A and the probabilities are in the inside of Table A.
- Read the  $z$  value off  $x.x$  from the row index and  $0.0x$  from the column index.
- Drawing pictures helps!

70

## Standard Normal Distribution

### Examples

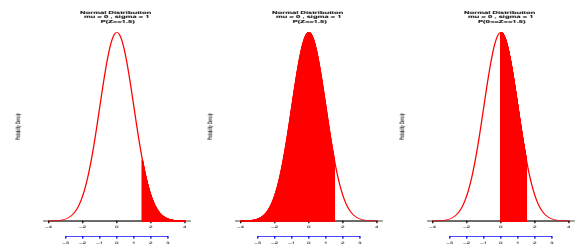
$$P(Z \geq 1.5) = 0.0668$$

$$P(Z \leq 1.5) = 1 - P(Z \geq 1.5)$$

$$= 1 - 0.0668 = 0.9332$$

$$P(0 \leq Z \leq 1.5) = P(Z \geq 0) - P(Z \geq 1.5)$$

$$= 0.5 - 0.0668 = 0.4332$$

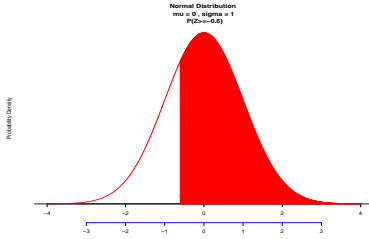


71

## Standard Normal Distribution

### More examples

$$\begin{aligned} P(Z \geq -0.6) &= 1 - P(Z \leq -0.6) \\ &= 1 - P(Z \geq 0.6) \\ &= 1 - 0.2743 = 0.7257 \end{aligned}$$

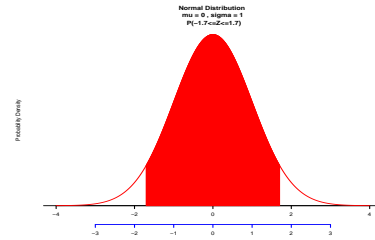


72

## Standard Normal Distribution

### More examples

$$\begin{aligned} P(-1.7 \leq Z \leq 1.7) &= 2 \times P(0 \leq Z \leq 1.7) \\ &= 2 \times (0.5 - 0.0446) = 0.9108 \end{aligned}$$

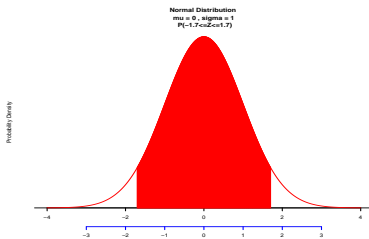


73

## Standard Normal Distribution

### More examples

$$\begin{aligned} P(-1.7 \leq Z \leq 1.7) &= 1 - P(Z \leq -1.7 \text{ or } Z \geq 1.7) \\ &= 1 - 2 \times P(Z \geq 1.7) \\ &= 1 - 2 \times 0.0446 = 1 - 0.0892 = 0.9108. \end{aligned}$$

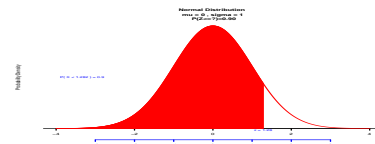


74

## Standard Normal Distribution

### Finding quantiles

What value  $z$  would give  $P(Z \leq z) = 0.90$ ?



- This  $z$  value would also give  $P(Z \geq z) = 1 - 0.90 = 0.10$ .
- Look up Table A for the  $z$  value that corresponds to 0.10.
- When  $z = 1.28$ ,  $P(Z \geq z) = 0.1003$ .
- When  $z = 1.29$ ,  $P(Z \geq z) = 0.0985$ .
- Thus  $z$  value sought after is a number between 1.28 and 1.29.
- More directly the table on page 92 of the bluebook gives  $z = 1.282$ .
- Interpretation: The  $z$  value sought after is the cut-off value where 90% of the population lies below (i.e. the 90<sup>th</sup> population quantile or the 90<sup>th</sup> population percentile).

75

## General Normal Distribution

### Definition and standardization

A r.v.  $Y$  is said to have a *normal distribution with parameters  $\mu$  and  $\sigma^2$*  and is written as:

$$Y \sim N(\mu, \sigma^2)$$

For example,  $Y \sim N(20, 3^2)$  stands for normal distribution with  $\mu = 20, \sigma^2 = 9$ .

**FACT:** Suppose  $Y \sim N(\mu, \sigma^2)$ , then a very useful transformation is

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

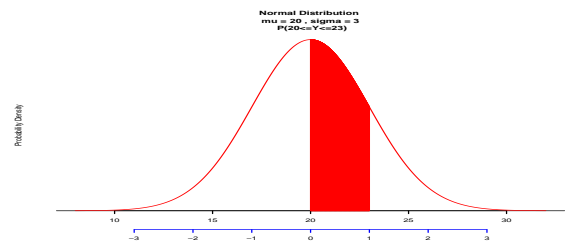
76

## General Normal Distribution

### Example

For  $Y \sim N(20, 3^2)$  (i.e.  $\mu = 20, \sigma^2 = 9$ ), we have

$$\begin{aligned} P(20 \leq Y \leq 23) &= P(20 - \mu \leq Y - \mu \leq 23 - \mu) \\ &= P\left(\frac{20 - \mu}{\sigma} \leq \frac{Y - \mu}{\sigma} \leq \frac{23 - \mu}{\sigma}\right) \\ &= P\left(\frac{20 - 20}{3} \leq Z \leq \frac{23 - 20}{3}\right) \\ &= P(0 \leq Z \leq 1) \\ &= P(Z \geq 0) - P(Z \geq 1) \\ &= 0.5 - 0.1587 = 0.3413 \end{aligned}$$



77

## General Normal Distribution

### Remarks

- Given a value  $y$ , the term

$$z = \frac{y - \mu}{\sigma}$$

is called the *z-score corresponding to  $y$* .

- It represents the number of standard deviations  $y$  is from  $\mu$ .
- For  $Y \sim N(20, 3^2)$ ,

23 is  $1\sigma$  away from  $\mu = 20$ .

20 is  $0\sigma$  away from  $\mu = 20$  (i.e. at  $\mu = 20$ ).

24 is  $\frac{4}{3}\sigma$  away from  $\mu = 20$ .

17 is  $-1\sigma$  away from  $\mu = 20$ .

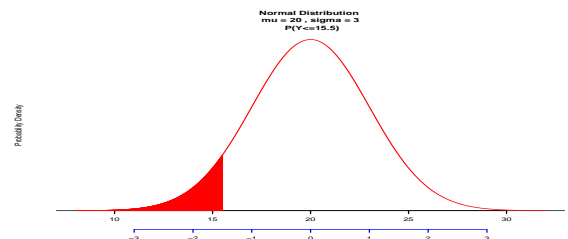
78

## General Normal Distribution

### More examples

Suppose  $Y \sim N(20, 3^2)$ , what is  $P(Y \leq 15.5)$ ?

$$\begin{aligned} P(Y \leq 15.5) &= P\left(\frac{Y - \mu}{\sigma} \leq \frac{15.5 - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{15.5 - 20}{3}\right) \\ &= P(Z \leq -1.5) \\ &= 0.0668 \end{aligned}$$



79

## General Normal Distribution

### Finding quantiles

Suppose  $Y \sim N(60, 10^2)$ , what value  $y^*$  would give

$$P(Y \leq y^*) = 0.90?$$

- From previous discussion, we know that

$$P(Z \leq 1.282) = 0.90$$

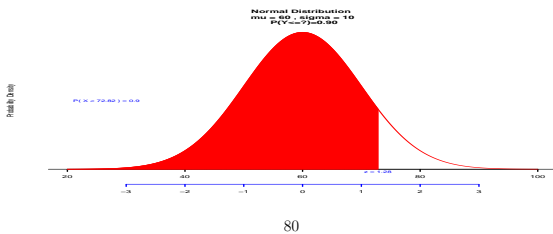
- Thus

$$P\left(\frac{Y - \mu}{\sigma} \leq 1.282\right) = 0.90$$

$$P(Y - \mu \leq 1.282\sigma) = 0.90$$

$$P(Y \leq \mu + 1.282\sigma) = 0.90$$

- Thus  $y^* = \mu + 1.282\sigma = 60 + 1.282 \times 10 = 72.82$ .
- $y^*$  is the 1.282 standard deviation above the mean.



## Binomial and Normal Distributions

### Key R commands

```
> # Y~B(3,0.5), compute P(Y=1)
> dbinom(x=1,size=3,prob=0.5)
[1] 0.375
> # Y~B(7,0.6), compute P(Y=3)
> dbinom(x=3,size=7,prob=0.6)
[1] 0.193536
> # Y~B(3,1/6), compute P(Y>0)
> pbinom(q=0,size=3,prob=1/6,lower.tail=F)
[1] 0.4212963
> # of by complement
> 1-dbinom(x=0,size=3,prob=1/6)
[1] 0.4212963
> # Z~N(0,1), compute P(Z>=1.5), P(Z<=1.5), P(0<=Z<=1.5),
> pnorm(1.5, lower.tail=F)
[1] 0.0668072
> pnorm(1.5, lower.tail=T)
[1] 0.9331928
> pnorm(0, lower.tail=F)-pnorm(1.5, lower.tail=F)
[1] 0.4331928
> # Z~N(0,1), compute P(Z>=-0.6)
> pnorm(-0.6, lower.tail=F)
[1] 0.7257469
> # Z~N(0,1), compute P(-1.7<=Z<=1.7)
> 1-2*pnorm(1.7, lower.tail=F)
[1] 0.910869
> # Z~N(0,1), compute P(Z<=?)=0.90
> qnorm(0.9, lower.tail=T)
[1] 1.281552
> # Y~N(20,9), compute P(20<=Y<=23), P(Y<=15.5)
> pnorm(q=20,mean=20,sd=3,lower.tail=F)-pnorm(23,mean=20,sd=3,lower.tail=F)
[1] 0.3413447
> pnorm(q=15.5,mean=20,sd=3,lower.tail=T)
[1] 0.0668072
> # Y~N(60,100), compute P(Y<=?)=0.90
> qnorm(p=0.9,mean=60,sd=10,lower.tail=T)
[1] 72.81552
```

81

## Sampling Distributions

### Overview

- Recall that a random variable (r.v.) is a variable that depends on the outcome of a chance situation.
- The probability distribution of a r.v. is described by the probability of all possible outcomes of the r.v.
- Now we put data and probability models together by viewing an observation (obs) as an outcome of a r.v.
- We consider  $n$  obs as outcomes of  $n$  r.v.'s.

82

## Sampling Distributions

### Definitions

- The r.v.'s  $Y_1, Y_2, \dots, Y_n$  (forming the sample data) are called a *random sample of size  $n$  from a given distribution* if

- the  $Y_i$ 's are independent r.v.'s;
- the  $Y_i$ 's are identically distributed (i.e. every  $Y_i$  has the same probability distribution).

- Conditions (1) and (2) are sometimes abbreviated to "the  $Y_i$ 's are iid".

- Further assume: (3) For each  $i$ ,  $E(Y_i) = \mu$ ,  $Var(Y_i) = \sigma^2$ .
- A *statistic* is any quantity whose value can be calculated from sample data. It is a r.v.

- For example, *sample mean* is a statistic:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$$

- Then we have under (1)–(3):

$$E(\bar{Y}) = \mu, \quad Var(\bar{Y}) = \frac{\sigma^2}{n}$$

- As  $n$  increases,  $Var(\bar{Y})$  decreases.

83

## Sampling Distributions

### Sample mean

- By (2) and (3),

$$\begin{aligned} E(\bar{Y}) &= E\left(\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right) \\ &= \frac{1}{n}E(Y_1 + Y_2 + \cdots + Y_n) \\ &= \frac{1}{n}(E(Y_1) + E(Y_2) + \cdots + E(Y_n)) \\ &= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu \end{aligned}$$

- By (1), (2), and (3),

$$\begin{aligned} \text{Var}(\bar{Y}) &= \text{Var}\left(\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right) \\ &= \frac{1}{n^2}\text{Var}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \frac{1}{n^2}(\text{Var}(Y_1) + \text{Var}(Y_2) + \cdots + \text{Var}(Y_n)) \\ &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) \\ &= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

84

## Sampling Distributions

### FACT 1

Suppose  $Y_1, Y_2, \dots, Y_n$  is a random sample from  $N(\mu, \sigma^2)$  distribution. Then

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

### Remarks

- For example, if  $Y_1, Y_2, \dots, Y_{40}$  are iid  $N(3, 4)$ , then

$$\bar{Y} \sim N\left(3, \frac{4}{40}\right) = N(3, 0.1)$$

- We may write  $\bar{Y} \sim N(\mu, \sigma_{\bar{Y}}^2)$ , where the variance of  $\bar{Y}$  is

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$$

and the corresponding standard deviation of  $\bar{Y}$  is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

(also known as the *standard error of the mean*).

85

## Sampling Distributions

### FACT 2

Suppose  $Y_1, Y_2, \dots, Y_n$  is a random sample from  $N(\mu, \sigma^2)$  distribution. Then

$$\sum_{i=1}^n Y_i \sim N(n\mu, n\sigma^2)$$

### Remarks

- For example, if  $Y_1, Y_2, \dots, Y_{40}$  are iid  $N(3, 4)$ , then

$$\sum_{i=1}^{40} Y_i \sim N(120, 160)$$

- Think of  $\sum_{i=1}^n Y_i = n\bar{Y}$ :

$$E\left(\sum_{i=1}^n Y_i\right) = n\mu, \quad \text{Var}\left(\sum_{i=1}^n Y_i\right) = n\sigma^2.$$

86

## Sampling Distributions

### FACT 3

Suppose  $Y_1, Y_2, \dots, Y_n$  is a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . If  $n$  is large, then the distribution of  $\bar{Y}$  is closely **approximated** by

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger  $n$  is, the better the approximation. This fact is also known as the *central limit theorem (CLT)* (for the sample mean).

87

## Sampling Distributions

### Example: fair die game

- Recall the fair die game example:  $Y =$  winning \$ with probability distribution

$$p(0) = 1/2, p(9) = 1/6, p(15) = 1/3$$

and mean  $\mu = 6.5, \sigma^2 = 46.25$ .

- If we repeat the game 1000 times independently, then for  $Y_1, Y_2, \dots, Y_{1000}$ , we have

$$E(\bar{Y}) = 6.5, \quad \text{Var}(\bar{Y}) = \frac{46.25}{1000}$$

- By CLT,  $\bar{Y}$  is **approximately**

$$N\left(6.5, \frac{46.25}{1000}\right).$$

- For example,

$$\begin{aligned} P(\bar{Y} \geq 7) &\approx P\left(Z \geq \frac{7 - 6.5}{\sqrt{\frac{46.25}{1000}}}\right) \\ &= P(Z \geq 2.32) = 0.0102. \end{aligned}$$

88

## Sampling Distributions

### FACT 4

Suppose  $Y_1, Y_2, \dots, Y_n$  is a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . If  $n$  is large, then the distribution of  $\sum_{i=1}^n Y_i$  is closely **approximated** by

$$\sum_{i=1}^n Y_i \sim N(n\mu, n\sigma^2)$$

The larger  $n$  is, the better the approximation. This fact is also known as the *central limit theorem (CLT)* (for the sample sum).

### Example: fair die game

- Note that  $n = 1000, E(\sum Y_i) = 1000 \times 6.5 = 6500,$   
 $\text{Var}(\sum Y_i) = 1000 \times 46.25 = 46250.$

- By FACT 4,  $\sum Y_i$  is **approximately**  
 $N(6500, 46250).$

- Then,

$$\begin{aligned} P(\sum Y_i \geq 7000) &\approx P\left(Z \geq \frac{7000 - 6500}{\sqrt{46250}}\right) \\ &= P(Z \geq 2.32) = 0.0102. \end{aligned}$$

89

## Sampling Distributions

### A quick summary

- Random sample, statistic, sample mean.
- If  $Y_i$  iid  $D(\mu, \sigma^2), E(\bar{Y}) = \mu, \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$
- If  $Y_i$  iid  $N(\mu, \sigma^2), \bar{Y} \sim N(\mu, \frac{\sigma^2}{n}).$
- If  $Y_i$  iid  $D(\mu, \sigma^2),$  approximately  $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$  (CLT).
- Similarly for  $\sum Y_i$  with mean  $n\mu$  and variance  $n\sigma^2.$

90

## Sampling Distributions

### Remarks on CLT

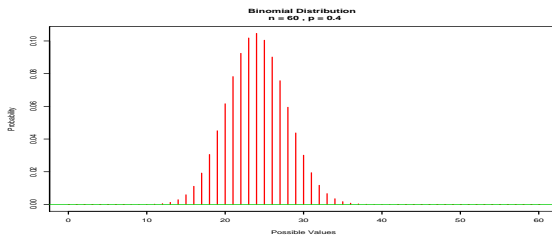
- The CLT is a remarkable result. It says that regardless of the shape of the original distribution, the taking of averages (or sums) results in normal distribution. Hence we only need to know the population mean and standard deviation to have an approximate distribution of  $\bar{Y}$  (or  $\sum Y_i$ ).
- Why is normal distribution so useful? Many natural phenomena give rise to normal variables; many variables can be transformed to normal distribution; and CLT.
- Simulations.
- How big must  $n$  be? Usually when  $n \geq 30,$  the CLT works well. But if  $Y_i$  has a highly skewed distribution, then we need larger  $n,$  in which case we may consider transformation to a symmetric distribution first.
- An application of the CLT is to approximate a binomial distribution by a normal distribution.

91

## Sampling Distributions

### Normal approximation of binomial

- Recall that if  $Y \sim B(n, p)$ , then  $\mu = np, \sigma^2 = npq$  where  $q = 1 - p$ .
- Let  $Y \sim B(60, 0.4)$ . Thus  $n = 60, p = 0.4$ , and  $\mu_Y = 24, \sigma_Y^2 = 14.4$ .
- Compute  $P(Y \leq 20) = P(Y = 0) + P(Y = 1) + \dots + P(Y = 20)$ ?



92

## Sampling Distributions

### Normal approximation of binomial

- Now, think of  $Y$  as the **sum** of  $n = 60$  iid r.v.'s, each of which is a 0/1 variable.
- Let  $Y_i = 1$  for a success with probability  $p$  and  $Y_i = 0$  for a failure with probability  $q$ . Then  $Y = \sum Y_i$  is the number of successes in  $n = 60$  trials.
- By the CLT on sum (FACT 4), we can approximate the distribution of  $Y$  by normal distribution with mean

$$\mu_Y = np = 24$$

and variance

$$\sigma_Y^2 = npq = 14.4 = (3.795)^2.$$

- Hence if  $Y \sim B(60, 0.4)$ , then  $Y_{\text{NA}} \sim N(24, 14.4)$ .

93

## Sampling Distributions

### Normal approximation of binomial

- Now, we approximate  $P(Y \leq 20)$  by

$$\begin{aligned} P(Y \leq 20) &\approx P(Y_{\text{NA}} \leq 20) \\ &= P\left(\frac{Y_{\text{NA}} - 24}{3.795} \leq \frac{20 - 24}{3.795}\right) \\ &= P(Z \leq -1.054) = 0.1460. \end{aligned}$$

- The exact probability is  $P(Y \leq 20) = 0.1786$

```
> pbinom(q=20, size=60, prob=0.4, lower.tail=T)
> [1] 0.1785702
```

- The discrepancy is largely due to the fact that we are approximating a discrete r.v. by a continuous r.v.
- We can make the normal approximation better by using continuity correction (cc).

$$\begin{aligned} P(Y \leq 20) &\approx P(Y_{\text{NA}} \leq 20.5) \\ &= P\left(\frac{Y_{\text{NA}} - 24}{3.795} \leq \frac{20.5 - 24}{3.795}\right) \\ &= P(Z \leq -0.922) = 0.1783. \end{aligned}$$

94

## Sampling Distributions

### Remarks

- Normal approximation is OK if  $np \geq 5$  and  $nq \geq 5$ . Always check this before using normal distribution to approximate binomial distribution.
- In the example,  $Y \sim B(60, 0.4)$ ,  $np = 24, nq = 36$ , thus it is OK to use normal approximation.

- Consider proportion of success (denoted as  $W$  before)

$$\hat{p} = \frac{Y}{n}$$

- For  $Y \sim B(60, 0.4)$ ,

$$E(\hat{p}) = p = 0.4, \quad \text{Var}(\hat{p}) = \frac{pq}{n} = \frac{0.4 \times 0.6}{60} = 0.004.$$

- Hence by CLT on average (FACT 3),

$$\hat{p}_{\text{NA}} \sim N(0.4, (0.0632)^2).$$

- Thus for the example  $Y \sim B(60, 0.4)$ ,

$$\begin{aligned} P(\hat{p} \leq 0.48) &\approx P(\hat{p}_{\text{NA}} \leq 0.48) \\ &= P\left(\frac{\hat{p}_{\text{NA}} - 0.4}{0.0632} \leq \frac{0.48 - 0.4}{0.0632}\right) \\ &= P(Z \leq 1.265) = 1 - 0.1038 = 0.8962. \end{aligned}$$

95

## Sampling Distributions

### Sample variance

Suppose  $Y_i$  are iid  $N(\mu, \sigma^2)$ . We know that the sample mean

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

What about the sample variance (also a statistic)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2?$$

### FACT

If  $Y_1, Y_2, \dots, Y_n$  form a random sample from  $N(\mu, \sigma^2)$ , then

$$V^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

where  $\chi_{n-1}^2$  is a *chi-squared distribution* with  $n-1$  degrees of freedom (d.f.).

## Sampling Distributions

### Sample variance

- Shape of the  $\chi_{n-1}^2$  distribution curve depends on the d.f.
- $V^2 \geq 0$
- The distribution curve is skewed (i.e. asymmetric).
- The mean and variance of  $V^2 \sim \chi_{n-1}^2$  are

$$E(V^2) = n-1, \quad \text{Var}(V^2) = 2(n-1).$$

- Think of  $V^2$  as a scaled version of  $S^2$ . Table B is in terms of  $V^2$ . Hence caution: always scale to  $V^2$  before using Table B.

- Since

$$S^2 = \frac{\sigma^2}{n-1} V^2$$

we have

$$E(S^2) = \sigma^2, \quad \text{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

- One reason to use  $(n-1)$  in  $S^2$  is to ensure  $E(S^2) = \sigma^2$ .
- As  $n$  increases,  $S^2$  estimates  $\sigma^2$  with more precision.

## Sampling Distributions

### Example

Let  $Y_1, Y_2, \dots, Y_7$  be a random sample from  $N(6, 3^2)$  (i.e.  $n = 7$ ). What is  $P(S^2 > 21)$ ?

$$\begin{aligned} P(S^2 > 21) &= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1) \times 21}{\sigma^2}\right) \\ &= P(V^2 > \frac{6 \times 21}{9}) \\ &= P(V^2 > 14) \end{aligned}$$

where  $V^2 \sim \chi_6^2$ . From Table B, we have

$$P(V^2 > 12.59) = 0.05, \quad P(V^2 > 14.45) = 0.025.$$

So

$$0.025 < P(V^2 > 14) < 0.05$$

and

$$0.025 < P(S^2 > 21) < 0.05$$

In R,

```
> pchisq(q=14, df=6, lower.tail=F)
> [1] 0.02963616
```

## Sampling Distributions

### Remarks

- In the example, if  $Y_i \sim N(200, 3^2)$ , then  $P(S^2 > 21)$  is the same, because it does not depend on  $\mu$ .
- Again caution: rescale  $S^2$  to  $V^2$  before using Table B.
- More on the d.f. for sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Consider  $n = 3$  and all the deviations. Once we know that

$$Y_1 - \bar{Y} = 3, \quad Y_2 - \bar{Y} = -1,$$

it follows that

$$Y_3 - \bar{Y} = -2.$$

Hence the complete information about deviations is contained in  $(n-1)$  of them.

- In general, d.f. represents the amount or pieces of information available about the spread.
- We use  $\bar{Y}, S^2$  instead of median and IQR, because they are easier to deal with.