

1. Recall B - binary, I - independent, N - n constant, S - same p.

- (a)  $X_1 \sim \text{Bin}(150, .8)$ .
- (b) Fails to have a constant number  $n$  of trials.
- (c) Not all trials have the same probability  $p$  of success.
- (d)  $X_4 \sim \text{Bin}(5000, 250/5000)$ .
- (e) If we think of each trial's possible outcomes as "is AB" or "isn't AB", and if we assume that all the progeny's genotypes are independent (biologically I think that this is approximately but not completely justified) then  $X_5 \sim \text{Bin}(28, 1/4)$ .

2.  $X \sim \text{Bin}(4, 0.42)$ .

(a)  $P(X = 0) = \binom{4}{0} 0.42^0 (1 - 0.42)^{4-0}$ :

```
> choose(4, 0) * 0.42^0 * (1 - 0.42)^(4 - 0)
[1] 0.1131650
```

(b)  $P(X = 2) = \binom{4}{2} 0.42^2 (1 - 0.42)^{4-2}$ :

```
> choose(4, 2) * 0.42^2 * (1 - 0.42)^(4 - 2)
[1] 0.3560458
```

(c)  $P(0 \leq X \leq 2) = \sum_{k=0}^2 \binom{4}{k} 0.42^k (1 - 0.42)^{4-k}$ :

```
> sum(lapply(0:2, function(k) choose(4, k) * 0.42^k * (1 - 0.42)^(4 -
+ k)))
[1] 0.7969989
```

(d)  $P(0 < X \leq 2) = \sum_{k=1}^2 \binom{4}{k} 0.42^k (1 - 0.42)^{4-k}$ :

```
> sum(lapply(1:2, function(k) choose(4, k) * 0.42^k * (1 - 0.42)^(4 -
+ k)))
[1] 0.6838339
```

3.  $p = 0.3$ .

(a) Note that  $n\hat{p}$  is  $\text{Bin}(n, p)$ , so

$$P(\hat{p} = q) = P(n\hat{p} = nq) = f_{\text{Bin}(n,p)}(nq)$$

where  $f_{\text{Bin}(n,p)}$  is the Binomial pmf. We can explicitly write out the distribution for  $n = 2$  as follows:

```
> xtable(lapply(0:2, function(k) c(q = k/2, probability.that.hat.p.equals.q = choose(2,
+ k) * 0.3^k * (1 - 0.3)^(2 - k))))
```

	q	probability.that.hat.p.equals.q
1	0.00	0.49
2	0.50	0.42
3	1.00	0.09

(b) Note

- i.  $E(\hat{p}) = (1/n)E(n\hat{p}) = (1/n)np = p$ .
- ii.  $\text{Var}(\hat{p}) = (1/n^2)\text{Var}(n\hat{p}) = (1/n^2)np(1-p) = (1/n)p(1-p)$
- iii.  $SE(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{(1/n)p(1-p)}$ .

So for  $n = 20$ ,

```
> (E.hat.p <- 0.3)
[1] 0.3
> (Var.hat.p <- (1/20) * 0.3 * (1 - 0.3))
```

```
[1] 0.0105
> (SE.hat.p <- sqrt(Var.hat.p))
[1] 0.1024695
```

(c) Note  $P(\hat{p} = 0.25) = P(n\hat{p} = n \cdot 0.25) = f_{\text{Bin}(20,0.3)}(20 \cdot 0.25)$ :

```
> (prob.hat.p.equals.0.25 <- (function() {
+   k <- 20 * 0.25
+   n <- 20
+   p <- 0.3
+   choose(n, k) * p^k * (1 - p)^(n - k)
+ })())
[1] 0.1788631
```

4. (a) We compute:

```
> (E.X <- (-4) * 0.1 + (-1) * 0.3 + 3 * 0.1 + 5 * 0.2 + 11 * 0.3)
[1] 3.9
> (E.XX <- (-4)^2 * 0.1 + (-1)^2 * 0.3 + 3^2 * 0.1 + 5^2 * 0.2 +
+ 11^2 * 0.3)
[1] 44.1
> (Var.X <- E.XX - E.X^2)
[1] 28.89
> (SD.X <- sqrt(Var.X))
[1] 5.374942
```

(b) Note that  $E(X) - SD(X) = 3.9 - 5.37 = -1.47$  and  $E(X) + SD(X) = 3.9 + 5.37 = 9.27$ . So the requested probability is  $P(X = -1) + P(X = 3) + P(X = 5) = 0.3 + 0.1 + 0.2 = 0.6$ .

(c) Note that  $E(X) - 2SD(X) = 3.9 - 2 * 5.37 = -6.84$  and  $E(X) + 2 * SD(X) = 3.9 + 2 * 5.37 = 14.64$ . So the requested probability is 1.

5. (a) Note  $E(3X_1 + 4X_2 - 2X_3) = 3EX_1 + 4EX_2 - 2EX_3 = 3 * 2 + 4 * 11 - 2 * (-3)$

```
> 3 * 2 + 4 * 11 - 2 * (-3)
[1] 56
```

(b) Note  $E(X_1 + 2X_1^2 - X_2 + X_2^2 - 3X_3^2) = EX_1 + 2E(X_1^2) - EX_2 + E(X_2^2) - 3E(X_3^2) = 2 + 2 * 20 - 11 + 140 - 3 * 31$

```
> 2 + 2 * 20 - 11 + 140 - 3 * 31
[1] 78
```

(c) Note  $\text{Var}(X_1) = E(X_1^2) - (EX_1)^2 = 20 - 2^2$ , etc.:

```
> (Var.X1 <- 20 - 2^2)
[1] 16
> (Var.X2 <- 140 - 11^2)
[1] 19
> (Var.X3 <- 31 - (-3)^2)
[1] 22
```

(d) Due to independence,  $\text{Var}(3X_1 + 4X_2 - 2X_3) = \text{Var}(3X_1) + \text{Var}(4X_2) + \text{Var}((-2)X_3) = 3^2\text{Var}(X_1) + 4^2\text{Var}(X_2) + (-2)^2\text{Var}(X_3)$ :

```
> 3^2 * Var.X1 + 4^2 * Var.X2 + (-2)^2 * Var.X3
[1] 536
```

(e) It is not possible, because  $\text{Var}(X_4) = E(X_4^2) - (EX_4)^2 = 12 - 4^2 = -4 < 0$  is not possible.

6.  $X \sim \text{Bin}(3, 0.6)$ .

(a) Using formula (1):

```
> sum(lapply(0:3, function(k) k * choose(3, k) * 0.6^k * (1 - 0.6)^(3 -
+ k)))
[1] 1.8
```

Using formula (3):

```
> 3 * 0.6
[1] 1.8
```

(b) Using formula (2):

```
> E.X <- sum(lapply(0:3, function(k) k * choose(3, k) * 0.6^k *
+ (1 - 0.6)^(3 - k)))
> E.XX <- sum(lapply(0:3, function(k) k^2 * choose(3, k) * 0.6^k *
+ (1 - 0.6)^(3 - k)))
> E.XX - E.X^2
[1] 0.72
```

Using formula (4):

```
> 3 * 0.6 * (1 - 0.6)
[1] 0.72
```

7. The researchers made a large number  $N$  of groups, each independent of the others. Call  $\hat{c}_i \sim \text{Bin}(n, p)$  the number of red alleles in group  $i$ , and  $\hat{p}_i = \hat{c}_i/n$  the proportion of red alleles in group  $i$ . Here  $n = 32$  and  $p = 0.5$ .

(a) On average, the mean proportion of red-eyed alleles among the groups is

$$\begin{aligned} E\left(\frac{1}{N} \sum_{i=1}^N \hat{p}_i\right) &= E\left(\frac{1}{Nn} \sum_{i=1}^N \hat{c}_i\right) \quad \text{definition of } \hat{p}_i \\ &= \frac{1}{Nn} \sum_{i=1}^N E\hat{c}_i \quad \text{linearity} \\ &= \frac{1}{Nn} \sum_{i=1}^N np \quad \text{since } c_i \sim \text{Bin}(n, p) \\ &= \frac{Nnp}{Nn} = p = 0.5 \end{aligned}$$

(b) On average, the standard deviation of the proportion of red-eyed alleles among the groups is

$$E\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{p}_i - \bar{p})^2}\right) = E\left(\sqrt{\frac{1}{N} \sum_{i=1}^N \hat{p}_i^2 - \bar{p}^2}\right)$$

where

$$\bar{p} = \frac{1}{N} \sum_{i=1}^N \hat{p}_i$$

(We could use  $N - 1$  in the denominator instead.)

I do not know a simple way to simplify this quantity further, due to the square root inside the expectation. Instead, I answer the question with (i) a simulation, and (ii) an analytical result for the variance.

(i) We simulate (10000 times) the proportion of red-eyed alleles for  $N$  groups, and we compute the average (across the simulations) standard deviation (among the groups) of the proportion. We do this for  $N = 10, 100, 1000$  and  $10000$  groups.

```

> Ns <- c(10, 100, 1000, 10000)
> expected.sds <- laply(Ns, function(N) {
+   sds <- laply(1:10000, function(iter) {
+     c.vec <- rbinom(N, size = 32, p = 0.5)
+     p.vec <- c.vec/32
+     sd(p.vec)
+   })
+   mean(sds)
+ })
> xtable(data.frame(N = Ns, expected.sd = expected.sds), digits = c(0,
+   0, 8))

```

	N	expected.sd
1	10	0.08591902
2	100	0.08818924
3	1000	0.08838141
4	10000	0.08837997

(ii) We answer the question analytically, substituting “variance” for “standard deviation”:

$$\begin{aligned}
E\left(\frac{1}{N} \sum_{i=1}^N (\hat{p}_i - \bar{p})^2\right) &= E\left(\frac{1}{N} \sum_{i=1}^N \hat{p}_i^2 - \left(\frac{1}{N} \sum_{i=1}^N \hat{p}_i\right)^2\right) \\
&= \left(\frac{1}{N} \sum_{i=1}^N E \hat{p}_i^2\right) - E\left(\left(\frac{1}{N} \sum_{i=1}^N \hat{p}_i\right)^2\right) \quad \text{linearity} \\
&= \left(\frac{1}{N} \sum_{i=1}^N E \hat{p}_i^2\right) - p \quad \text{part (a)} \\
&= \left(\frac{1}{Nn^2} \sum_{i=1}^N E \hat{c}_i^2\right) - p \quad \text{definition of } p_i \\
&= \left(\frac{1}{Nn^2} \sum_{i=1}^N np(1-p+np)\right) - p \quad (*) \\
&= \frac{Nnp(1-p+np)}{Nn^2} - p = \frac{p(1-p-n+np)}{n}
\end{aligned}$$

Above (\*) holds because the second moment (about 0, not the mean) of a  $\text{Bin}(n, p)$  random variable is  $np(1-p+np)$ .

(c) 40%.

(d) The probability that a randomly chosen group has exactly 50% red-eyed alleles is  $P(\hat{p}_i = 1/2) = P(\hat{c}_i = 16) = \binom{32}{16} 0.5^{16} (1-0.5)^{16}$ .

```
> choose(32, 16) * 0.5^16 * (1 - 0.5)^16
```

```
[1] 0.1399499
```

(e) The probability that a randomly chosen group will have 30 or more red-eyed alleles is  $P(\hat{c}_i \geq 30) = \sum_{k=30}^{32} \binom{32}{k} 0.5^k (1-0.5)^{32-k}$

```
> sum(laply(30:32, function(k) choose(32, k) * 0.5^k * (1 - 0.5)^(32 -
+   k)))
```

```
[1] 1.231674e-07
```