Chapter 8

Sampling Distributions

Sampling distributions are probability distributions of statistics.

8.1 Distribution of the Sample Mean

Sampling distribution for random sample average, $\bar{X}$, is described in this section. The *central limit theorem (CLT)* tells us *no matter what the original parent distribution*, sampling distribution of $\bar{X}$ is typically normal when $n \geq 30$. Related to this,

$$\mu_{\bar{X}} = \mu_X, \quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}, \quad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}.$$

Exercise 8.1 (Distribution of the Sample Mean)

1. *Practice with CLT: average, $\bar{X}$.*

   (a) *Number of burgers.*
   Number of burgers, X, made per minute at Best Burger averages $\mu_X = 2.7$ burgers with a standard deviation of $\sigma_X = 0.64$ of a burger. Consider average number of burgers made over random $n = 35$ minutes during day.

   i. $\mu_{\bar{X}} = \mu_X =$ (circle one) **2.7** / **2.8** / **2.9**.

   ii. $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{0.64}{\sqrt{35}} =$ **0.10817975** / **0.1110032** / **0.13099923**.

   iii. $P\left(\bar{X} > 2.75\right) \approx$ (circle one) **0.30** / **0.32** / **0.35**.
   (Stat, Calculators, Normal, Mean: 2.7, Std. Dev.: 0.10817975, Prob(X $\geq$ 2.75) = [?], Compute.)

   iv. $P\left(2.65 < \bar{X} < 2.75\right) \approx$ (circle one) **0.36** / **0.39** / **0.45**.
   (Stat, Calculators, Normal, Between, Mean: 2.7, Std. Dev.: 0.10817975, Prob(2.65 $\leq$ X $\leq$ 2.75))

   (b) *Temperatures.*
   Temperature, X, on any given day during winter in Laporte averages $\mu_X =$

0 degrees with standard deviation of $\sigma_X = 1$ degree. Consider average temperature over random $n = 40$ days during winter.

i. $\mu_{\bar{X}} = \mu_X =$ (circle one) **0** / **1** / **2**.

ii. $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1}{\sqrt{40}} =$ **0.0900234** / **0.15811388** / **0.23198455**.

iii. $P\left(\bar{X} > 0.2\right) \approx$ (circle one) **0.03** / **0.10** / **0.15**.
(Stat, Calculators, Normal, Mean: 0, Std. Dev.: 0.15811388,
Prob(X > 0.2) = [?], Compute.)

iv. $P\left(\bar{X} > 0.3\right) \approx$ (circle one) **0.03** / **0.10** / **0.15**.
(Stat, Calculators, Normal, Mean: 0, Std. Dev.: 0.15811388,
Prob(X > 0.3) = [?], Compute.)
Since $P\left(\bar{X} > 0.3\right) \approx 0.03 < 0.05$, 0.3^o **is** / **is not** unusual.

(c) *Another example.*
Suppose X has distribution where $\mu_X = 1.7$ and $\sigma_X = 1.5$.

i. $\mu_{\bar{X}} = \mu_X =$ (circle one) **2.3** / **1.7** / **2.4**.

ii. $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.5}{\sqrt{49}} =$ **0.0243892** / **0.14444398** / **0.21428572**.

iii. If $n = 49$, $P(-2 < \bar{X} < 2.75) \approx$ (circle one) **0.58** / **0.86** / **0.999**.
(Stat, Calculators, Normal, Between, Mean: 1.7, Std. Dev.: 0.21428572, Prob(-2 ≤ X ≤ 2.75))

iv. **True** / **False**.
If $n = 15$, $P(-2 < \bar{X} < 2.75)$ cannot be calculated since $n = 15 < 30$.

v. $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.5}{\sqrt{15}} =$ **0.0243892** / **0.14444398** / **0.38729835**.

vi. If $n = 15$ *and* normal, $P(-2 < \bar{X} < 2.75) \approx$ **0.75** / **0.78** / **0.997**.
(Stat, Calculators, Normal, Between, Mean: 1.7, Std. Dev.: 0.38729835, Prob(-2 ≤ X ≤ 2.75))

(d) *Dice average.*
What is the chance, in $n = 30$ rolls of a fair die, average is between 3.3 and 3.7, $P\left(3.3 < \bar{X} < 3.7\right)$? What if $n = 3$?

i. $\mu_{\bar{X}} = \mu_X = 1\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) =$ (circle one) **2.3** / **3.5** / **4.3**.
(Blank data table. Relabel var1 die, var2 frequency. Type 1, 2, 3, 4, 5, 6, in die column, and 1, 1, 1, 1, 1, 1 in frequency column. Data, Save Data, 8.1 die distribution. Dat, compute expression, Expression: frequency/6, New column name: P(x), Compute. Stat, Calculators, Custom, Values in: die, Weights in: P(x), Okay. Notice, Mean: 3, Std. Dev.: 1.7078252.)

ii. $\sigma_X = \sqrt{(1-3.5)^2\left(\frac{1}{6}\right) + \cdots + (6-3.5)^2\left(\frac{1}{6}\right)} =$
1.7078252 / **2.131145** / **3.3409334**.

iii. If $n = 30$, $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.71}{\sqrt{30}} =$ (circle one) **0.31** / **0.75** / **1.14**.

iv. If $n = 30$, $P\left(3.3 < \bar{X} < 3.7\right) \approx$ (circle one) **0** / **0.20** / **0.48**.
(Stat, Calculators, Normal, Between, Mean: 3.5, Std. Dev.: 0.3118048, Prob(3.3 ≤ X ≤ 3.7))

v. **True** / **False**.
If $n = 3$, $P\left(3.3 < \bar{X} < 3.7\right)$ cannot be calculated because $n = 3 < 30$.

2. *Understanding CLT: Montana fishing trip.*

   (a) *Sampling distributions of average, $n = 1, 2, 3$.*

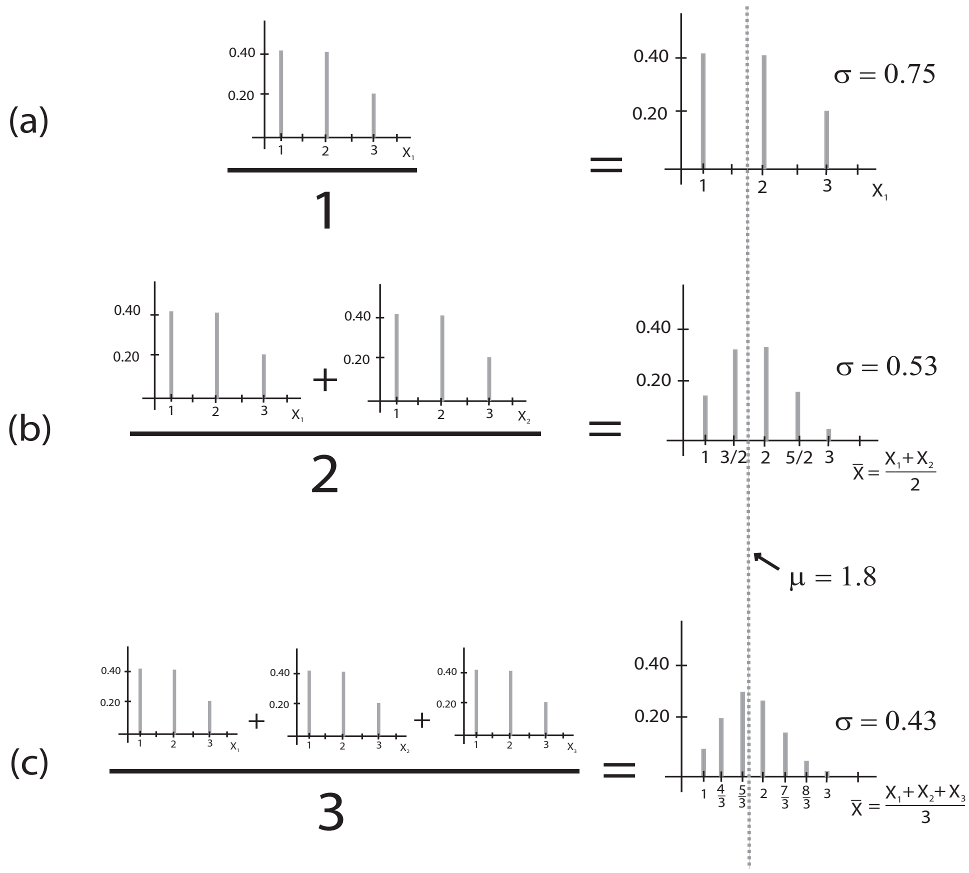

Figure 8.1 (Comparing sampling distributions of sample mean)

As random sample size, n, increases, sampling distribution of average, $\bar{X}$, changes shape and becomes more (circle one)

i. rectangular–shaped.
ii. bell–shaped.
iii. triangular–shaped.

Central limit theorem (CLT) says *no matter what the original parent distribution*, sampling distribution of average is typically normal when $n > 30$.

(b) In addition to sampling distribution becoming more normal–shaped as random sample size increases, mean of average, $\mu_{\bar{X}} = 1.8$ (circle one)

i. decreases and is equal to $\frac{\sigma_X^2}{n}$,
ii. remains same and is equal to $\mu_X = 1.8$,
iii. increases and is equal to $n\mu_X$,

and standard deviation of average, $\sigma_{\bar{X}}$ (circle one)

i. decreases and is equal to $\frac{\sigma_X}{\sqrt{n}}$.
ii. remains same and is equal to σ_X.
iii. increases and is equal to $n\sigma_X$.

(c) After $n = 30$ trips to lake, sampling distribution in average number of fish caught is essentially *normal* (why?) where
$\mu_{\bar{X}} = \mu_X =$ (circle one) **1.2** / **1.5** / **1.8**,
$\sigma_{\bar{X}} = \frac{0.75}{\sqrt{30}} \approx$ **0.12677313** / **0.13693064** / **0.2449987**,
(Data, Compute expression, Expression: 0.75/sqrt(30), Compute.)
and chance average number of fish is *less than* 1.95 is
$P(\bar{X} < 1.95) \approx$ (circle one) **0.73** / **0.86** / **0.94**.
(Stat, Calculators, Normal, Mean: 1.8, Std. Dev.: 0.13693064,
Prob(X < 1.95) = ?, Compute.)

(d) After $n = 35$ trips to lake, sampling distribution in average number of fish caught is essentially normal where
$\mu_{\bar{X}} = \mu_X =$ (circle one) **1.2** / **1.5** / **1.8**,
$\sigma_{\bar{X}} = \frac{0.75}{\sqrt{35}} \approx$ **0.12677313** / **0.13693064** / **0.2449987**,
(Data, Compute expression, Expression: 0.75/sqrt(35), Compute.)
and chance average number of fish is *less than* 1.95 is
$P(\bar{X} < 1.95) \approx$ (circle one) **0.73** / **0.88** / **0.94**.
(Stat, Calculators, Normal, Mean: 1.8, Std. Dev.: 0.12677313,
Prob(X < 1.95) = ?, Compute.)

(e) Chance average number of fish is less than 1.95 after *30 trips*, $P(\bar{X} < 1.95) \approx 0.86$, is **smaller than** / **larger than** chance average number of fish is less than 1.95 after *35 trips*, $P(\bar{X} < 1.95) \approx 0.88$.

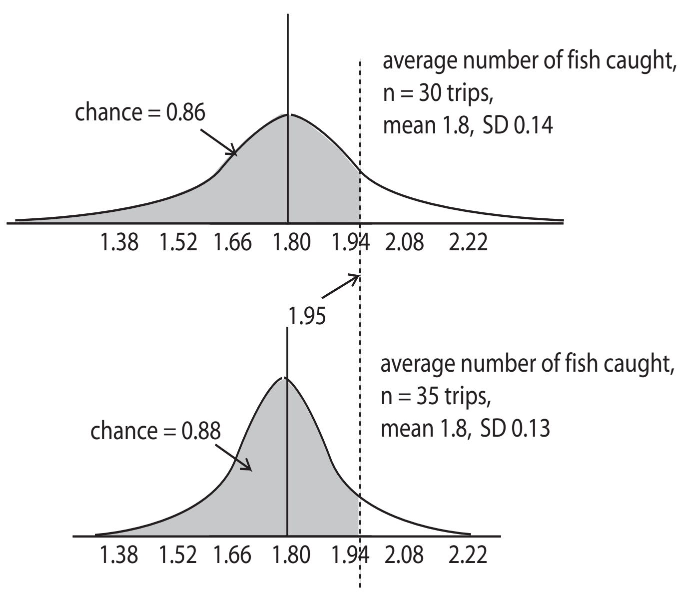

Figure 8.2 (Chance when $n = 30$ compared to chance when $n = 35$)

(f) The CLT is useful because (circle *one or more*):

i. No matter what original parent distribution is, as long as a large enough random sample is taken, average of this sample follows a normal distribution.

ii. In practical situations where it is not known what parent probability distribution to use, as long as a large enough random sample is taken, average of this sample follows a normal distribution.

iii. Rather than having to deal with many different probability distributions, as long as a large enough random sample is taken, average of this sample follows *one* distribution, normal distribution.

iv. Many distributions in statistics rely in one way or another on normal distribution because of CLT.

(g) **True** / **False** Central limit theorem requires not only $n \geq 30$, but also a *random sample* of size $n \geq 30$ is used.

3. *Simulating sampling distribution for average*[1]*: Montana fishing trip.*

[1]Text uses simulations, rather than exact sampling distributions, in discussion of CLT.

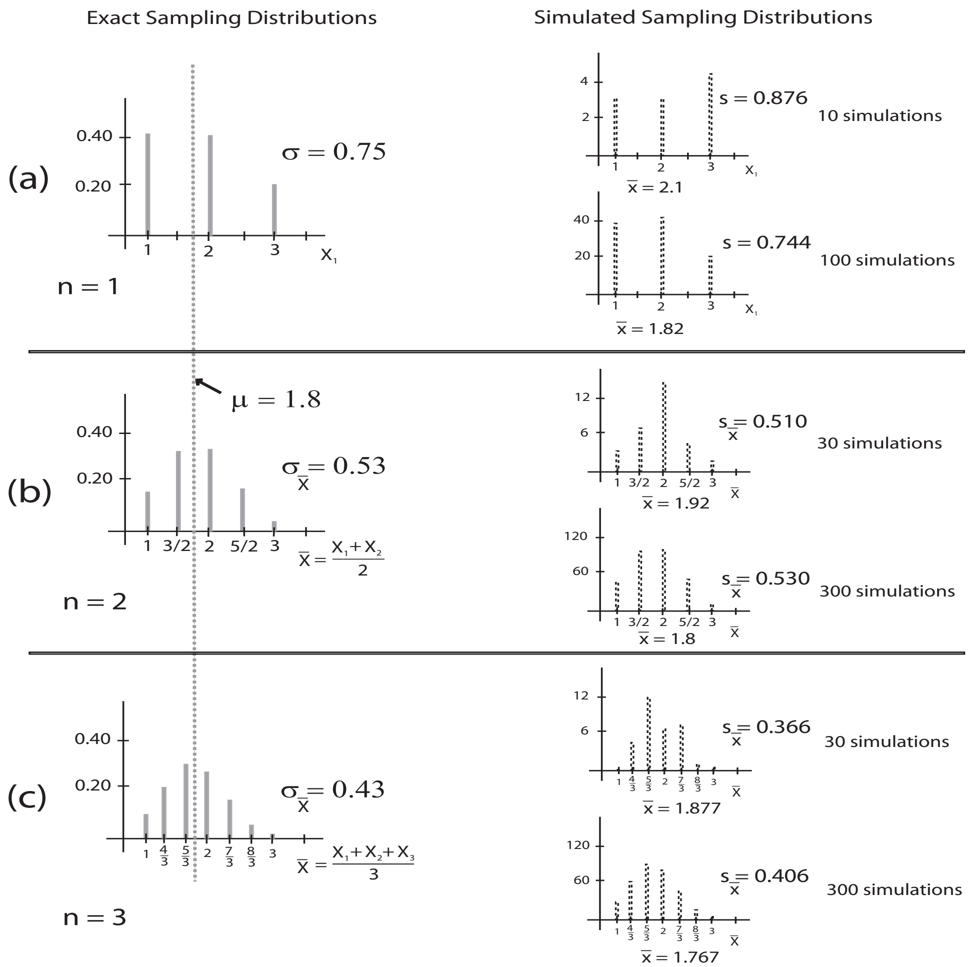

Figure 8.3 (Simulating sampling distributions: Montana fishing trip.)

(a) *Figure 8.4(a). Parent distribution,* $n = 1$.
Approximate simulated histogram shape and $(\bar{x}, s)$ all (choose one)
improve, are typically closer to
worsen, are typically farther away from
parent distribution shape and $(\mu, \sigma) \approx (1.8, 0.75)$.
as number of simulations increases, from 10 to 100.

(b) *Figure 8.4(b). Sampling distribution average,* $n = 2$.
Approximate simulated histogram shape and $(\bar{x}, s)$ all (choose one)
improve, are typically closer to
worsen, are typically farther away from
parent distribution shape and $(\mu, \sigma) \approx (1.8, 0.53)$.
as number of simulations increases, from 30 to 300.

(c) *Figure 8.4(c). Sampling distribution average,* $n = 3$.
Approximate simulated histogram shape and $(\bar{x}, s)$ all (choose one)

improve, are typically closer to
worsen, are typically farther away from
parent distribution shape and $(\mu, \sigma) \approx (1.8, 0.43)$.
as number of simulations increases, from 10 to 100.

(d) *Sample Size Versus Number of Simulations.*
As number trips to lake (*sample size*) increases, $n = 1$ to $n = 3$, sampling distribution of average **does** / **does not** become more normal.
As number of *simulations* increase, approximate sampling distribution **does** / **does not** become more normal *unless* distribution normal.

8.2 Distribution of the Sample Proportion

Central limit theorem (CLT) tells us no matter what the original parent distribution, sampling distribution of random[2] sample proportion[3], $\hat{p} = \frac{X}{n}$, is typically normal when $np(1-p) \geq 10$ and $n \leq 0.05N$. Related to this,

$$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Exercise 8.2 (Distribution of the Sample Proportion)

1. *Proportion of wins.*
Lawyer estimates she wins 40% of her cases ($p = 0.4$), and currently represents $n = 50$ defendants. Let X represent number of wins (of 50 cases) and so $\hat{p} = \frac{X}{n}$ proportion of wins (of 50 cases). Use CLT to approximate chance she wins at least one–half of her cases,

$$P\left(\hat{p} > \frac{1}{2}\right) = P\left(\hat{p} > 0.5\right).$$

(a) *Check assumptions.*
Since $np(1-p) = 50(0.4)(1-0.4) = 12 \geq 10$, assumptions necessary to proceed with approximation are (choose one) **satisfied** / **violated**.

(b) $\mu_{\hat{p}} = p =$ (circle one) **0.3** / **0.4** / **0.5**.

(c) $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{50}} \approx$ **0.0011655** / **0.03855667** / **0.06928203**.
(Data, Compute expression, Expression: sqrt(0.4*0.6/50), Compute.)

[2]If population finite, simple random sample size must be small, no more than 5% of population size, $\frac{n}{N} \leq 0.05$, to ensure independence of items in sample.

[3]Previously discussed normal approximation to binomial is an example of CLT. We look at this approximation again, only focus on sample proportion, $\hat{p} = \frac{X}{n}$, rather than sample number, X.

(d) $P(\hat{p} > 0.5) \approx$ (circle one) **0.07** / **0.11** / **0.13**.
(Stat, Calculators, Normal, Mean: 0.4, Std. Dev.: 0.06928203,
Prob(X ≥ 0.5) = [?], Compute.)
Since $P(\hat{p} > 0.5) \approx 0.07 > 0.05$, $\hat{p} = 0.5$ is **typical** / **unusual**.

2. *Another example.*
Let $p = 0.63$ and $n = 45$ and approximate

$$P(\hat{p} < 0.41).$$

(a) *Check assumptions.*
Since $np(1-p) = 45(0.63)(1-0.63) = 10.4895 > 10$, assumptions necessary to proceed with approximation are (choose one) **satisfied** / **violated**.

(b) $\mu_{\hat{p}} = p =$ (circle one) **0.54** / **0.60** / **0.63**.

(c) $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.63(1-0.63)}{45}} \approx$ **0.06177628** / **0.07197221** / **0.08900453**.
(Data, Compute expression, Expression: sqrt(0.63*0.37/45), Compute.)

(d) $P(\hat{p} < 0.41) \approx$ (circle one) **0.001** / **0.002** / **0.003**.
(Stat, Calculators, Normal, Mean: 0.63, Std. Dev.: 0.07197221,
Prob(X ≤ 0.41) = [?], Compute.)
Since $P(\hat{p} < 0.41) \approx 0.001 < 0.05$, $\hat{p} = 0.63$ is **typical** / **unusual**.

3. *And another example.*
Let $p = 0.25$ and $n = 42$, $N = 10000$ and approximate

$$P(0.22 < \hat{p} < 0.28).$$

(a) *Check assumptions.*
Since $np(1-p) = 42(0.25)(0.75) = 7.875 < 10$ and $n = 42 < 0.05N = 0.05(10000) = 500$, assumptions necessary to proceed with approximation are **satisfied** / **violated**. So we won't.

4. *Proportion of wins again.*
Lawyer estimates she wins 40% of her cases ($p = 0.4$), and currently represents $n = 50$ defendants. What is the chance[4] she wins at least 30 cases?

(a) $\mu_{\hat{p}} = p =$ (circle one) **0.3** / **0.4** / **0.5**.

(b) $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{50}} \approx$ **0.00123234** / **0.0387786** / **0.06928203**.
(Data, Compute expression, Expression: sqrt(0.4*0.6/50), Compute.)

(c) $P(X \geq 30) = P\left(\hat{p} \geq \frac{30}{50}\right) \approx$ (circle one) **0.001** / **0.002** / **0.003**.
(Stat, Calculators, Normal, Mean: 0.4, Std. Dev.: 0.06928203,
Prob(X ≥ 0.6) = [?], Compute.)

[4] Ignore, do not use, the continuity correction here.