Chapter 9

Statistical inference and sampling distributions

Recall from Chapter 1

Statistical inference: drawing conclusions about a population based on information obtained from a sample (e.g., estimating the mean annual household income of families in Australia from a survey).

Descriptive measures of a population are called ***parameters***, while descriptive measures calculated from a sample are called ***statistics***.

In many applications of statistical inference, we draw **conclusions about a parameter of a population by using a sample statistic**.

Introduction to sampling distributions

In real life, calculating the *parameters* of populations is prohibitive because populations are very large.

Rather than investigating the whole population, we take a sample, calculate a *statistic* related to the *parameter* of interest and make an inference.

The ***sampling distribution*** of the *statistic* is the tool that tells us how close the statistic is to the parameter.

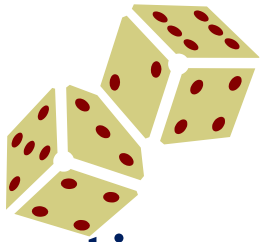
Sampling Distributions

A sampling distribution is created by, as the name suggests, ***sampling***.

The method we will employ to derive the sampling distribution uses the ***rules of probability*** and the ***laws of expected value and variance***.

For example, consider the roll of one and two dice…

Sampling distribution of the sample mean



Sampling distribution of a single die

A fair **die** is thrown **infinitely many times**, with the random variable X = Number of spots showing on any throw. **This is our (large!) population**

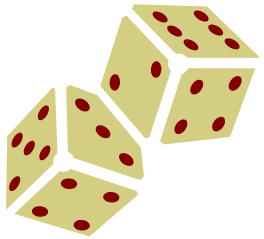
The probability distribution of X is:

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

But **suppose you don't know this.**

and the mean and variance are calculated as follows:

Sampling distribution of the sample mean…



Sampling distribution of a single die

The mean and variance are calculated as:

$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

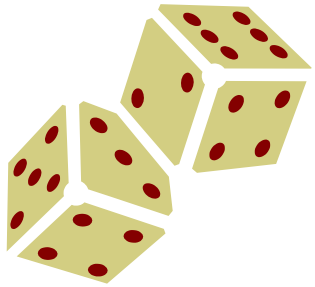
$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

These are **population parameters**

Sampling Distribution of Two Dice

A sampling distribution of the sample mean $\bar{X}$ is created by looking at all samples of size n=2 (i.e. two dice) and their means…



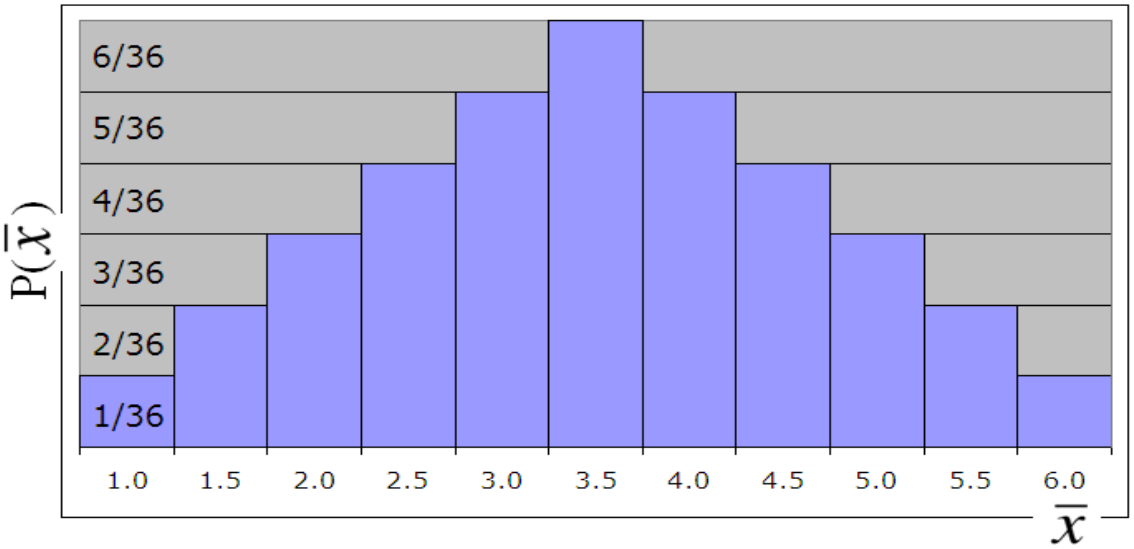
Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

While there are 36 possible samples of size 2, there are only 11 values for $\bar{X}$, and some (e.g. $\bar{X}$ =3.5) occur more frequently than others (e.g. $\bar{X}$ =1.0).

Sampling Distribution of Two Dice

The ***sampling distribution*** of $\bar{X}$ is shown below:

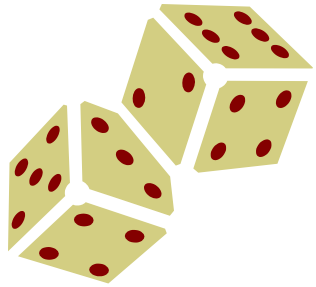
$\bar{x}$	$p(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

$$\sigma_{\bar{x}}^2 = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

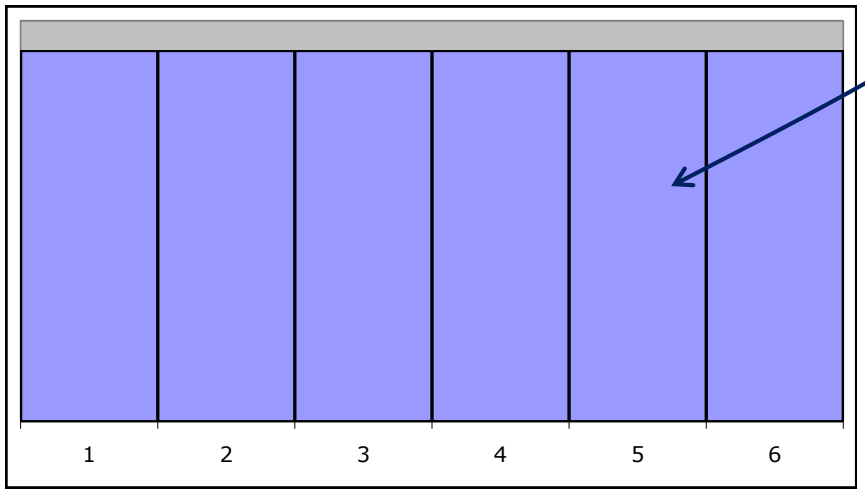
$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{1.46} = 1.21$$

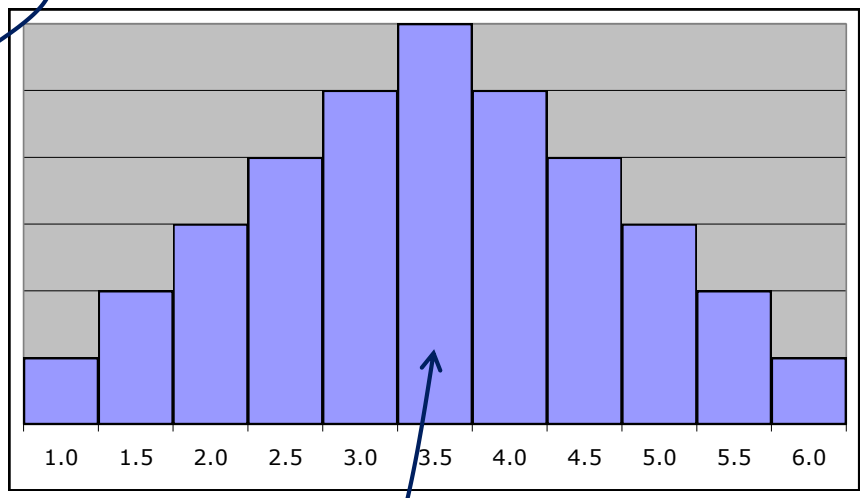


This is the distribution of a sample statistic (the mean)

Compare…

Compare the distribution of X…





…with the sampling distribution of $\bar{x}$.

As well, note that: $\mu_{\bar{x}} = \mu$

$$\sigma_{\bar{x}}^2 = \sigma^2 / 2$$

Generalize…

We can generalize the mean and variance of the sampling of two dice:

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \sigma^2/2$$

…to **n**-dice:

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

The standard deviation of the sampling distribution of the sample mean is called the ***standard error of the mean.***

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

Central Limit Theorem

Intuitive fact: if a random sample is drawn from a normal population, then the sampling distribution of the sample mean $\bar{X}$ is normally distributed for all values of *n* (sample size).

Theorem: if a random sample is drawn from **any** population, the sampling distribution of the sample mean $\bar{X}$ is approximately normal for a sufficiently large sample size (usually $n \geq 30$).

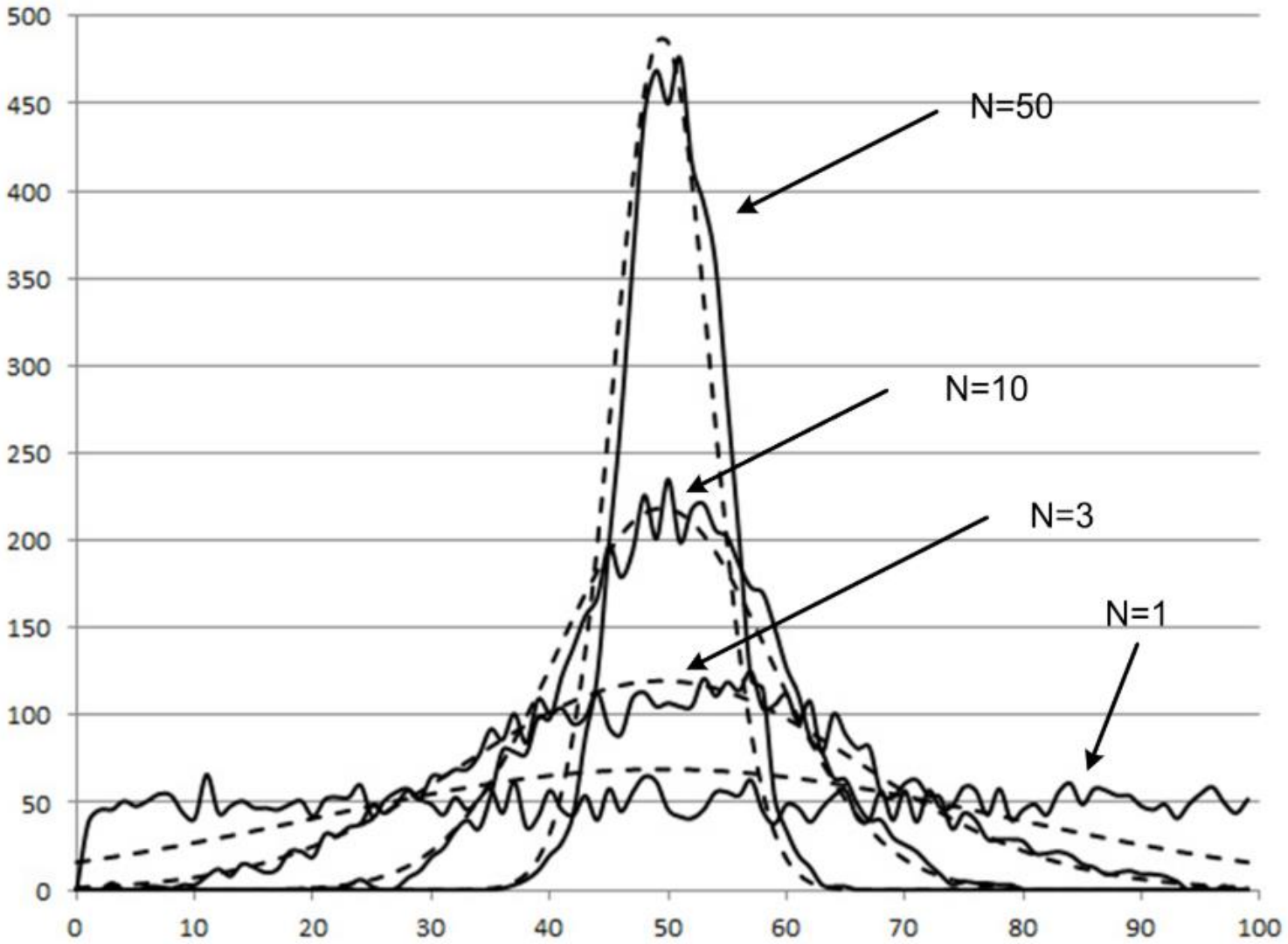
This theorem is known as the ***Central Limit Theorem (CLT)***.

Central Limit Theorem…

The larger the sample size, the more closely the sampling distribution of $\bar{X}$ will resemble a normal distribution.

In most practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of $\bar{X}$.

Central Limit Theorem…



© Douglas Brooks

Central Limit Theorem…

"Statistics is pretty powerful. For example, it allows me to take a sample and then say things like:

I know (with 95% confidence) how close my sample mean is to the true population mean, even though I have no idea what the true mean is. Or even what the population looks like! I know how close I am to the truth without knowing what the truth is!

And if that's not close enough, I know what to do to get a closer estimate. That seems powerful to me.

It is the Central Limit Theorem that allows me to do this"

(Douglas Brooks)

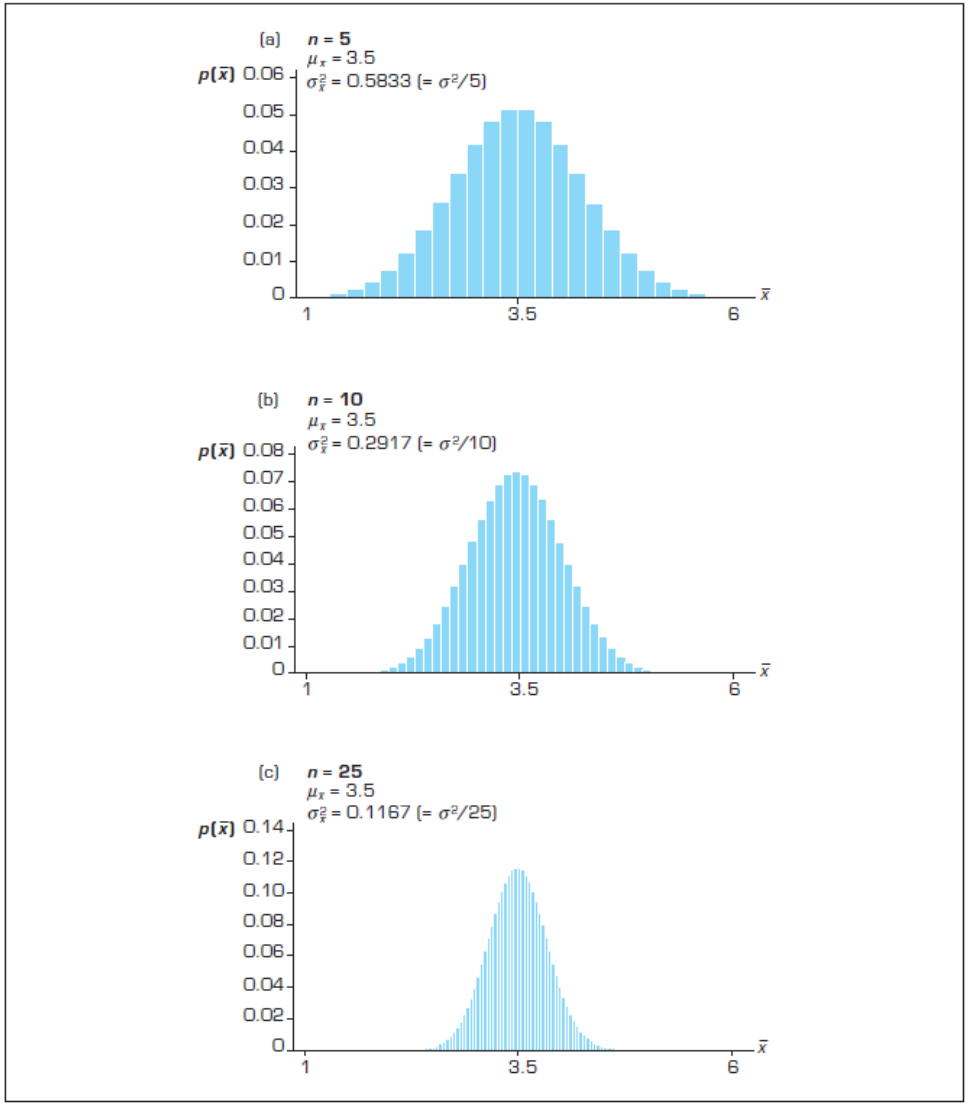
Sampling Distribution of the Sample Mean: summary of key results

1. $\mu_{\bar{x}} = \mu_x$

2. $\sigma^2_{\bar{x}} = \dfrac{\sigma^2_x}{n}$

3. *If x is* normal, $\bar{x}$ is normal. If x is non-normal $\bar{x}$ is approximately normally distributed for sufficiently large sample size ($n \geq 30$).

Figure 9.3 Sampling distributions of $\bar{X}$ when $n = 5$, 10 and 25



Sampling Distribution of the Sample Mean

We can standardise the sampling distribution of the sample mean $\bar{X}$ as

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Sampling Distribution of the Sample Mean

The summaries above assume that the population is infinitely large. However, if the population is finite, the standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where *N* is the population size and

$$\sqrt{\frac{N-n}{N-1}}$$

is the ***finite population correction factor***.

Sampling Distribution of the Sample Mean

If the population size is large relative to the sample size the finite population correction factor is close to 1 and can be ignored.

We will treat any population that is at least 20 times larger than the sample size as large.

In practice, most applications involve populations that qualify as large.

As a consequence the finite population correction factor is usually omitted.

Example 1

The weight of each '32g' chocolate bar is normally distributed with a mean of 32.2 g and a standard deviation of 0.3 g.

i. Find the probability that, if a customer buys one chocolate bar, that bar will weigh more than 32 g.

ii. Find the probability that, if a customer buys a pack of 4 bars, the mean weight of the bars will be more than 32 g.

Example 1: Solution

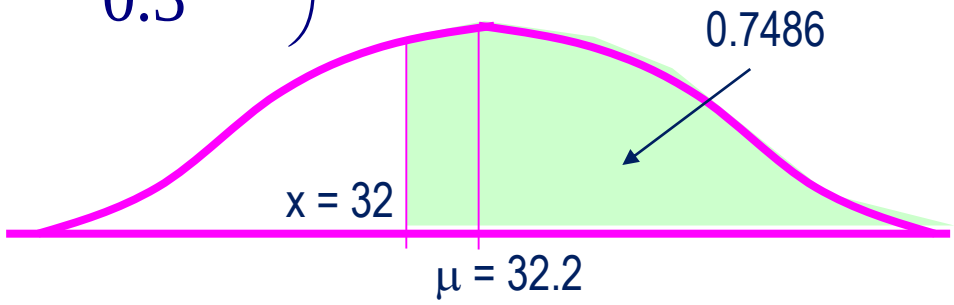
i. Find the probability that, if a customer buys one chocolate bar, that bar will weigh more than 32 g.

The random variable X is the weight of a chocolate bar which is normally distributed with $\mu_X = 32.2$, $\sigma_X = 0.3$. Therefore,

$$P(x > 32) = = P\left(\frac{x - \mu}{\sigma_x} > \frac{32 - 32.2}{0.3}\right)$$

$$= P(Z > -0.67)$$

$$= 0.7486$$



Example 1: Solution

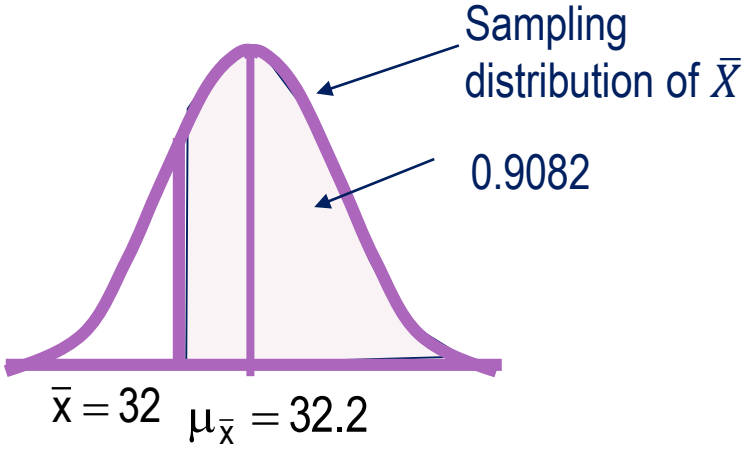
ii. Find the probability that, if a customer buys a pack of 4 bars, the mean weight of the 4 bars will be more than 32 g.

The random variable of interest is the mean weight per chocolate bar, $\bar{X}$. The population X is normally distributed with $\mu_X = 32.2$, $\sigma_X = 0.3$.

Therefore, $\bar{X}$ is normal with $\mu_{\bar{X}} = 32.2$, $\sigma_{\bar{X}} = 0.3/\sqrt{4}$.

$$P(\bar{x} > 32) = P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} > \frac{32 - 32.2}{0.3/\sqrt{4}}\right)$$

$$= P(z > -1.33) = 0.9082$$

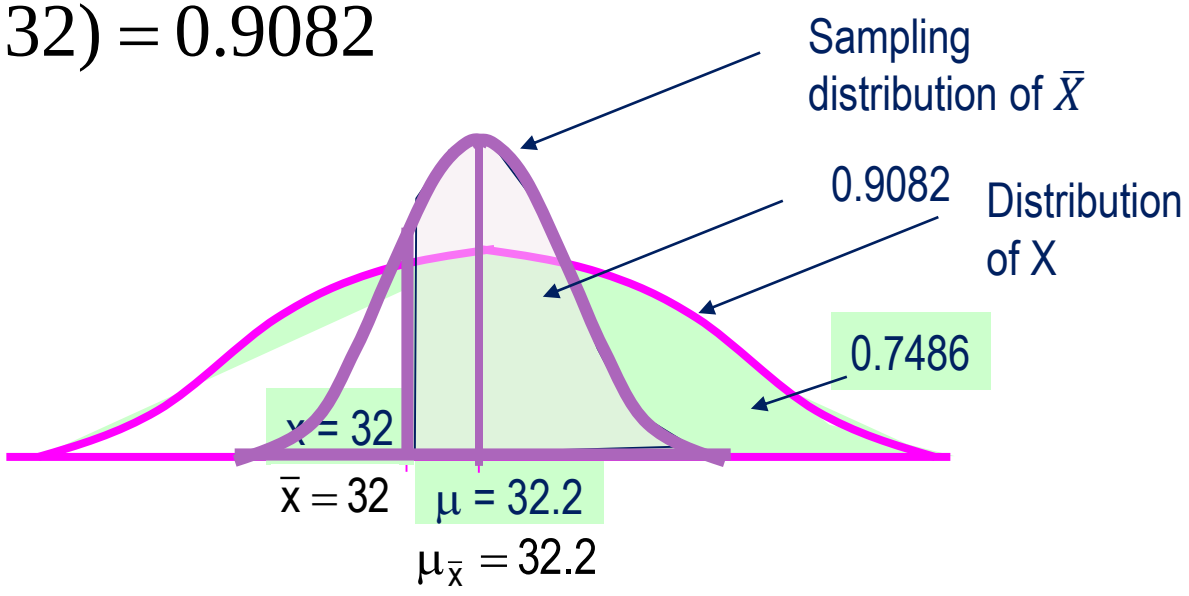


Example 1: Solution

Graphical illustration of the two probabilities:

i. $P(X > 32) = 0.7486$

ii. $P(\bar{X} > 32) = 0.9082$



Example 2

The average weekly income of graduates one year after graduation is $600. Suppose the distribution of weekly income is normally distributed with standard deviation of $100.

i. What is the probability that 25 randomly-selected graduates have an average weekly income of less than $550?

ii. If a random sample of 25 graduates actually had an average weekly income of $550, what would you conclude about the validity of the claim that the average weekly income is $600?

Example 2: Solution

i. Let X be the weekly income of graduates one year after graduation. X is normally distributed with $\mu_X = 600$, $\sigma_X = 100$ and n = 25.

Therefore, $\mu_{\bar{X}} = 600$, $\sigma_{\bar{X}} = 100/\sqrt{25}$ and $\bar{X}$ is normally distributed. We want to find $P(\bar{X} < 550)$.

$$P(\bar{x} < 550) = P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{550 - 600}{100/\sqrt{25}}\right)$$
$$= P(Z < -2.5) = 0.0062$$

Example 2: Solution

ii. With μ = 600, the probability of having a sample mean of 550 is very low (0.0062). So the claim that the average weekly income is \$600 is probably unjustified.

It would be more reasonable to assume that μ is smaller than \$600, because then a sample mean of \$550 becomes more probable.

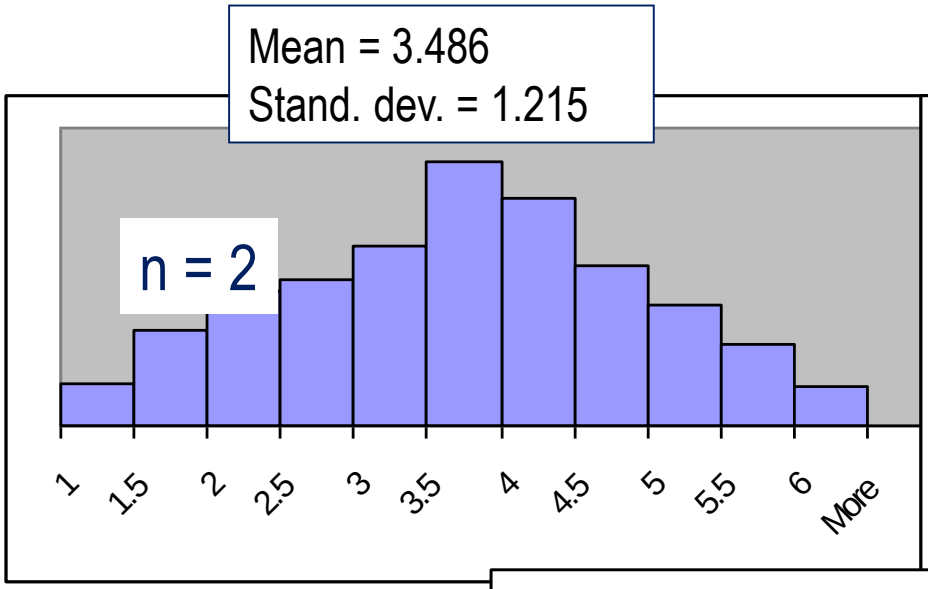
Creating the Sampling Distribution by Computer Simulation

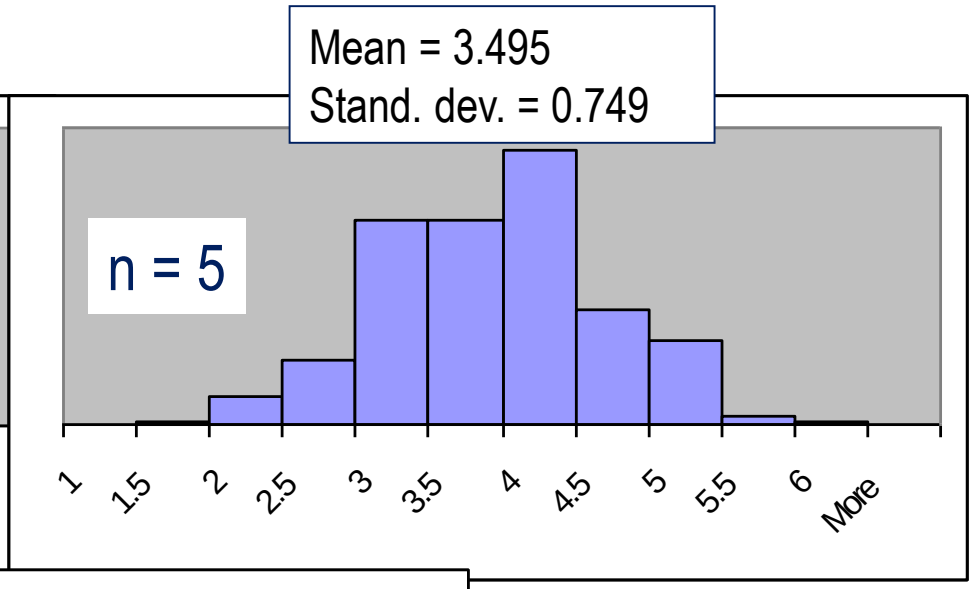
By producing data sets of random numbers that come from given distributions, we can verify probabilistic and statistical characteristics.

We simulate a dice-tossing experiment (creating the distribution of the average).

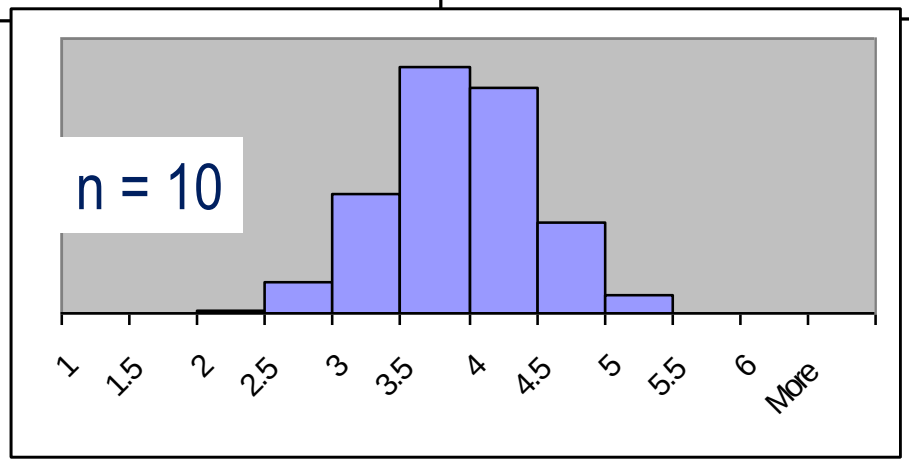
Effects of an increasing sample size on the distribution of the mean are shown.

Simulation of Dice Tossing





Mean = 3.494
Stand. dev. = 0.544



Excel
Creating a simulated distribution of the mean

Using samples of size 2 calculate the sample means

Calculate the means.

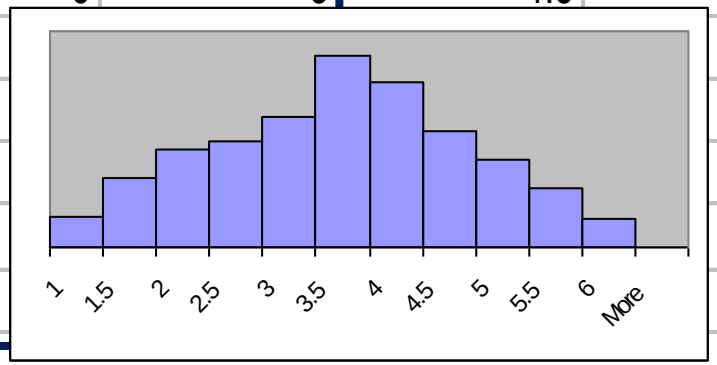
Type the bin.

Create samples of size two.

Create a histogram for the distribution of the mean.

Observation1	Observation2	Sample Mean	Bin
4	6	5	1
6	6	6	1.5
1	3	2	2
6	1	3.5	2.5
1	1	1.5	3
1	1	1	3.5
2	1	1.5	4
3	3	2.5	4.5
3	3	3.5	5
3	3	3	5.5
6	3	4.5	6

Bin	*Frequency*
1	28
1.5	65
2	90
2.5	98
3	121
3.5	177
4	152
4.5	107
5	81
5.5	55
6	26
More	0



9.5 Sampling distribution of the sample proportion

The parameter of interest for nominal data is the *proportion of times* a particular outcome (success) occurs.

Let X be the number of times a particular outcome (success) occurs in n repeated trials.

To estimate the population proportion of successes, p, we use the sample proportion $\hat{p}$ =X/n.

The sampling distribution of X is binomial.

We prefer to use normal approximation to the binomial distribution to make inferences about $\hat{p}$.

Sampling Distribution of a Sample Proportion $\hat{p}$

From the laws of expected value and variance, it can be shown that $E(\hat{p}) = p$ and $V(\hat{p}) = pq/n$

If both $np \geq 5$ and $nq \geq 5$, then

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

is approximately standard normal distributed.

Example 3

The Laurier company's brand has a market share of 30%. In a survey, 1000 consumers were asked which brand they prefer. What is the probability that more than 32% of the respondents say they prefer the Laurier brand?

Example 3: Solution

The number of respondents who prefer Laurier is binomial with $n = 1000$ and $p = 0.30$.

Also, $np = 1000(0.3) = 300 > 5$,

$nq = 1000(1 - 0.3) = 700 > 5$.

Therefore, $\hat{p}$ is normal with mean $p = 0.30$ and standard error

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.30(1-0.30)}{1000}} = 0.01449$$

Hence

$$P(\hat{p} > 0.32) = P\left(\frac{\hat{p} - p}{\sqrt{pq/n}} > \frac{0.32 - 0.30}{0.01449} \right) = 0.0838$$