What is a sampling distribution? Take a sample from a population and draw conclusions based on the statistics of the sample.

2 Types of Sampling Distributions:
Mean & Proportion

- A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population.

Goal of data analysis is to make statistical inferences, i.e., use statistics calculated from samples to estimate the values of population parameters. The Sample Mean is a statistic used to estimate the population mean (which is a parameter) and the Sample Proportion is a statistic used to estimate the population mean (also a parameter). Here we are trying to draw conclusions about a population, not about the sample! A Sampling Distribution represents the every possible sample that can occur from a population.

Unbiased
The arithmetic mean is said to be unbiased because the average of all possible sample means (of a given sample size n) will be equal to the population mean $\mu$. Based on the property of unbiasedness it is always true that $\mu_X = \mu$. 

Why Study Sampling Distributions

- Sample statistics are used to estimate population parameters
  - e.g.: $\bar{X} = 50$ Estimates the population mean $\mu$
- Problems: different samples provide different estimate
  - Large samples gives better estimate; Large samples costs more
  - How good is the estimate?
- Approach to solution: theoretical basis is sampling distribution

What is a sampling distribution? Take a sample from a population and draw conclusions based on the statistics of the sample.
The sample average is $\bar{X}$. It may be the case that $\bar{X}$ will be slightly different for each sample taken from the population (frame). The sample distribution of the mean is the histogram distribution of all possible sample means. Taking all possible samples will tend to remove and bias caused by extremes.

If all of the sample means are averaged, the mean of these values, $\mu_\bar{X}$, is equal to the population mean $\mu$. This says that the sample arithmetic mean is an *unbiased* estimator of the population mean.

\[ \mu_\bar{X} = \mu \]
Developing Sampling Distributions

- Assume there is a population ...
- Population size $N=4$
- Random variable, $X$, is age of individuals
- Values of $X$: 18, 20, 22, 24 measured in years

**Summary Measures for the Population Distribution**

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} = 2.236$$

Uniform Distribution
Developing Sampling Distributions

(continued)

<table>
<thead>
<tr>
<th>1st Observation</th>
<th>2nd Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>18.18</td>
</tr>
<tr>
<td>20</td>
<td>20.20</td>
</tr>
<tr>
<td>22</td>
<td>22.22</td>
</tr>
<tr>
<td>24</td>
<td>24.24</td>
</tr>
</tbody>
</table>

16 Samples Taken with Replacement

WITH REPLACEMENT

16 Sample Means

Each of these entries is the mean of the corresponding samples taken where n-2.
From all possible sample means of size \( n = 2 \) we create a histogram to derive the sample distribution.

Where the original population had a uniform distribution, equal chance of any particular age within the range, the SAMPLE DISTRIBUTION has a Normal distribution.

NO LONGER UNIFORM !!!
Developing Sampling Distributions

(continued)

Summary Measures of Sampling Distribution

\[ \mu_{\bar{X}} = \frac{\sum_{i=1}^{N} \bar{X}_i}{N} = \frac{18 + 19 + 19 + \ldots + 24}{16} = 21 \]

\[ \sigma_{\bar{X}} = \sqrt{\frac{\sum_{i=1}^{N} (\bar{X}_i - \mu_{\bar{X}})^2}{N}} \]

\[ = \sqrt{\frac{(18 - 21)^2 + (19 - 21)^2 + \ldots + (24 - 21)^2}{16}} = 1.58 \]

These are calculated using the SAMPLE MEANS DATA (16 data points).

For sampling with replacement:

As n increases, \( \sigma_{\bar{X}} \) decreases

Larger sample size

Smaller sample size
Comparing the Population with its Sampling Distribution

Original Population Distribution

Population
N = 4
μ = 21    σ = 2.236

Sample Means Distribution
n = 2
μ_\bar{x} = 21    σ_\bar{x} = 1.58

Sample Means Distribution

P(X)

\begin{align*}
A &= (18) \\
B &= (20) \\
C &= (22) \\
D &= (24)
\end{align*}

\begin{align*}
X &= 18, 19, 20, 21, 22, 23, 24
\end{align*}
As the sample size increases, the standard error of the mean decreases, so that a larger proportion of sample means are closer to the population mean. The sample means are distributed more tightly around the population mean as the sample size is increased.

How the arithmetic mean varies from sample to sample is expressed statistically by the value of the standard deviation of all possible sample means. This measure of variability in the mean from sample to sample is referred to as the Standard Error of the Mean, $\sigma_x$.
Unbiased: the average of all the possible sample means of a given sample size \( n \) will be equal to the population mean \( \mu \).

The Central Limit Theorem says that the sample means always distribute as Normal, regardless of the originating distribution.

Mean distribution varies less than the Median Distribution.

The Central Limit Theorem says that the sample means always distribute as Normal, regardless of the originating distribution.
Central Limit Theorem requires at least:

- $n = 30$ for ANY population distribution
- $n = 15$ for a symmetric population distribution
- $n = 1$ for a normal population distribution

But don’t necessarily stop at $n = 30$ because you may end up with a sample distribution which is normal but with a HUGE standard deviation.
If a population is normal with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

(This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

**Z-value for Sampling Distribution of the Mean**

Z-value for the sampling distribution of $X$:

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{n}}$$
As $n$ increases the distribution becomes normal according to the central limit theorem. Note that the means of the population and sampling distribution are the same.
Central Limit Theorem

As n increases

For Any Population Distribution

the sampling distribution becomes almost normal regardless of shape of population

\( \bar{X} \)

How Large is Large Enough?

- For most distributions, \( n > 30 \)
- For fairly symmetric distributions, \( n > 15 \)
- For normal distribution, the sampling distribution of the mean is always normally distribute \( \text{Even for } n = 1 \).

- We can apply the Central Limit Theorem:
  - Even if the population is not normal,
  - …sample means from the population will be approximately normal as long as the sample size is large enough.
Example: ALWAYS DRAW THE DISTRIBUTION

6.76 (pg 271)

An orange juice producer buys all his oranges from a large orange grove. The amount of juice squeezed from each of these oranges is approximately normally distributed with a mean of 4.70 ounces and a standard deviation of 0.40 ounce.

a. What is the probability that a randomly selected orange will contain between 4.70 and 5.00 ounces?

b. What is the probability that a randomly selected orange will contain between 5.00 and 5.50 ounces?

c. 77% of the oranges will contain at least how many ounces of juice?

\[ P_s \approx N(4.7, .4) \]

a) \[ P(4.7 \leq x \leq 5.0) \]

\[ z = \frac{5.0 - 4.7}{.4} = .75 \quad \Rightarrow \quad P(x \leq 5.0) = .7734 \quad \text{and} \quad z = \frac{4.7 - 4.7}{.4} = 0 \quad \Rightarrow \quad P(x \leq 4.7) = .5 \]

\[ .7734 - .5000 = .2734 = 27.34\% \]

b) \[ P(5.0 \leq x \leq 5.5) \]

\[ z = \frac{5.5 - 4.7}{.4} = 2 \quad \Rightarrow \quad P(x \leq 5.5) = .9772 \quad \text{and} \quad z = \frac{5.0 - 4.7}{.4} = .75 \quad \Rightarrow \quad P(x \leq 5.0) = .7734 \]

\[ .9772 - .7734 = .2038 = 20.38\% \]

d. Between what two values (in ounces) symmetrically distributed around the population mean will 80% of the oranges fall?

Suppose that a sample of 25 oranges is selected:

e. What is the probability that the sample mean will be at least 4.60 ounces?

f. Between what two values symmetrically distributed around the population mean will 70% of the sample means fall?

g. 77% of the sample means will be above what value?
c) ...contain “at least” how many ounces… This means we want to find the value of x which yields $1 - 77\% = 23\%$. We will cite the area greater than this x value as the region representing 77% of oranges containing “at least” x amount of juice.

$$P(x \leq ?) = 1 - 77\% = 23\%$$

In the Z table find the entry for 23% and read it’s Z value.

$$z = \frac{x - 4.7}{.4} = -.74 \quad \Rightarrow \quad x = -.74 \times .4 + 4.7 = 4.404$$

d) \quad P(x_1 \leq x \leq x_2) = 80\%

$$P(x_1 \leq x) = 10\%$$

$$z = \frac{x_1 - 4.7}{.4} = -1.28$$

$$x_1 = -1.28 \times .4 + 4.7 = 4.188$$

$$x_2 = 4.7 + (4.7 - 4.188) = 5.212$$

e) \quad$\text{Now we are working with sample distributions of sample size n=25.}$

What is the probability that the sample mean will be “at least” 4.6 ounces?

Here we are told the population distribution is normal so with our sample size of 25 from a symmetrically distributed population the central limit theorem holds. This means our sample mean will equal our population mean, $\mu_X = \mu$, and the sample standard error (variation) will be $\sigma_X = \frac{\sigma}{\sqrt{n}}$.

$$P(X \geq 4.6) = ?$$

$$z = \frac{4.6 - 4.7}{.4} = -1.25 \quad \Rightarrow \quad \frac{4.6 - 4.7}{\sqrt{25}}$$

$$P(4.6 \leq x) = 1 - .1056 = 89.44\%$$
f) \( P(x_1 \leq \bar{X} \leq x_2) = 70\% \)

\[
P(x_1 \leq \bar{X}) = 15\%
\]

\[
z = \frac{x_1 - 4.7}{0.4} = -1.04
\]

\[
x_1 = -1.04 \times 0.08 + 4.7 = 4.6168
\]

\[
x_2 = 4.7 + (4.7 - 4.6168) = 4.7832
\]

g) 77% of the sample means will be above what value?

\( P(x \geq x_i) = 77\% \)

In the Z table find the entry

for 1-77% = 23% and read it’s Z value.

\[
z = \frac{x_2 - 4.7}{0.4} = -0.74 \quad \Rightarrow
\]

\[
x = -0.74 \times \frac{0.4}{\sqrt{25}} + 4.7 = 4.6408
\]
Example

- Suppose a population has mean \( \mu = 8 \) and standard deviation \( \sigma = 3 \). Suppose a random sample of size \( n = 36 \) is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used \((n \geq 30)\)
- … so the sampling distribution of \( \bar{X} \) is approximately normal
- … with mean \( \mu_{\bar{X}} = 8 \)
- … and standard deviation \( \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5 \)

\[
P(7.8 < \mu_{\bar{X}} < 8.2) = P \left( \frac{7.8 - 8}{3/\sqrt{36}} < \frac{\mu_{\bar{X}} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}} \right)
= P(-0.5 < Z < 0.5) = 0.3830
\]
Example: \( \mu = 8 \quad \sigma = 2 \quad n = 25 \)

\[
P(7.8 < \bar{X} < 8.2) = \frac{7.8 - 8}{2/\sqrt{25}} < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{8.2 - 8}{2/\sqrt{25}}
\]

\[
P(7.8 < \bar{X} < 8.2) = P(\frac{-0.5}{0.4} < Z < \frac{0.5}{0.4}) = P(-1.25 < Z < 1.25) = 0.3830
\]

Central Limit Theorem requires at least:

- \( n \geq 30 \) for ANY population distribution
- \( n \geq 15 \) for a symmetric population distribution
- \( n = 1 \) for a normal population distribution

We don’t really know the population distribution in the above example but we do know it is symmetric (kind of) and we do know \( n > 15 \) so the central limit theorem will hold. (we know it is symmetric because it has a mean and a standard deviation.)
Population Proportions \( (p) \)

- **Categorical variable**
  - e.g.: Gender, voted for Bush, college degree

- **Proportion of population \( (p) \) having a characteristic**
  - Ex. \( p = \frac{52}{100} = 52\% \) <-- population

- **Sample proportion provides an estimate**
  - \( p_s = \frac{X}{n} = \frac{\text{number of successes}}{\text{sample size}} \) <-- \( p_s \) is a random variable

- **If two outcomes, \( X \) has a binomial distribution**
  - Possess or do not possess characteristic
  - \( p_s \) will vary among samples, it is a random variable.

Larger sample sizes will move \( p_s \) closer to \( p \).

**Notation:** we will write this slightly differently;

\[
P(50 \text{ out of 100 people voted for Bush when } p = .52) = P(p_s = .5 \text{ when } p = .52)
\]

Says probability sample proportion is .5 when population proportion is .52.

**Example:** \( P(50 \text{ out of 100 people voted for Bush when } p = .52) = ? \)

In this case \( p \) is the population proportion that voted for Bush.

Use Binomial Distribution to solve.

\[
P(p_s = .5 \text{ when } p = .52) = \text{Binomdist}(50, 100, .52, \text{False}) = 7.347\%
\]

In this example the sample probability can only be in certain increments such as \( p_s = 0, .01, .02, \ldots, .98, .99, 1 \). This is because your number of successes and your number of trials have to resolve to a whole number. Can’t have 1/3 of a person voting for Bush. This will effect the shape (resolution) of the distribution.

**Example:** \( (p_s \leq .6) = \text{binomdist}(60, 100, .52, \text{TRUE}) - \text{binomdist}(50, 100, .52, \text{FALSE}) = 64.79\% \)

Becomes .49 because problem calls for \( .5 \leq \), less than or **equal** meaning .5 is in the set we keep!
Sampling Distribution of Sample Proportion

Approximated by
normal distribution

Required to use normal approximation

\[
np \geq 5 \\
n(1 - p) \geq 5
\]

Mean:

\[
\mu_{p_s} = p
\]

Standard error:

\[
\sigma_{p_s} = \sqrt{\frac{p(1 - p)}{n}}
\]

This says that the mean of the sample proportion, \( p_s \), is the population proportion, \( p \).

\( np \geq 5 \) and \( n(1 - p) \geq 5 \) are required to use the normal approximation for categorical / binomial data. Will have trouble meeting this requirement if \( n \) is very small or \( p \) is either very small or very large.

When the conditions are met we can use the normal approximation to find the probabilities. However, the normal approximations will still be a little different than the binomial solution (binomial solution is more accurate).

Under these conditions the mean of the sample proportion, \( \mu_{p_s} \), is the population proportion, \( p \).

Example: find \( P(0.5 \leq p_s \leq 0.6 \) when \( p = 0.52 \) using the normal approximation.

\[
np = (100)(0.52) = 52 \quad \text{and} \quad n(1 - p) = (100)(0.48) = 48
\]

From last page, binomial solution = 64.79%. Now solve using the normal approximation.

\[
\text{normdist}(0.5, 0.52, \sqrt{\frac{0.52(1 - 0.52)}{100}}, \text{TRUE}) = 0.94534 \quad \text{(binomial was 0.95606)}
\]

\[
\text{normdist}(0.5, 0.52, \sqrt{\frac{0.52(1 - 0.52)}{100}}, \text{TRUE}) = 0.34446 \quad \text{(binomial was 0.30815)}
\]

\[
P(0.5 \leq p_s \leq 0.6 \) when \( p = 0.52 \)) = 0.94534 - 0.34446 = 60.088\% \text{ (binomial: 64.79\%)}
\]
**Standardizing Sampling Distribution of Proportion**

\[
Z \approx \frac{p_s - \mu_{p_s}}{\sigma_{p_s}} = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}
\]

**Sampling Distribution**

\[\sigma_{p_s}\]

\[\mu_{p_s}\]  \[p_s\]  \[\mu_Z = 0\]  \[Z\]

**Standardized Normal Distribution**

\[\sigma_Z = 1\]

**Standardize** \(p_s\) **to a Z value with the formula:**

\[
Z = \frac{p_s - p}{\sigma_{p_s}} = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}
\]

**If sampling is without replacement** and \(n\) is greater than \(5\%\) of the population size, then \(\sigma_{p_s}\) must use the **finite population correction factor:**

\[
\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}
\]
Example:

6.81 (pg 272)

DiGiorno’s frozen pizza has some of the most creative and likeable advertisements on television. USA Today’s Ad Track claims that 20% of viewers like the ads “a lot” (Theresa Howard, “DiGiorno Campaign Delivers Major Sales,” www.usatoday.com, April 1, 2002). Suppose that samples of 400 television viewers are shown the advertisements.

a. What proportion of the samples will have between 18% and 22% who like the ads “a lot”?

b. What proportion of the samples will have between 16% and 24% who like the ads “a lot”?

c. What proportion of the samples will have between 14% and 26% who like the ads “a lot”?

d. What proportion of the samples will have between 12% and 28% who like the ads “a lot”?

\[
np = (400)(.2) = 80 \quad \text{and} \quad n(1-p) = (400)(.8) = 320
\]

Both are \(\geq 5\) so we can use the normal approx.

\[
\mu_p = p \quad \sigma_p = \sqrt{\frac{p(1-p)}{n}}
\]

a) \(P(.18 \leq p_s \leq .22)

\[
z = \frac{.22 - .2}{.02} = 1 \quad \Rightarrow \quad P(p_s \leq .22) = .8413
\]

\[
z = \frac{.18 - .2}{.02} = -1 \quad \Rightarrow \quad P(p_s \leq .18) = .1587
\]

Notice the range asked for is equal to \(\pm \sigma\) so we already knew the answer = 68.27%.

b) \(P(.16 \leq p_s \leq .24)

\[
z = \frac{.24 - .2}{.02} = 2 \quad \Rightarrow \quad P(p_s \leq .24) = .9772
\]

\[
z = \frac{.16 - .2}{.02} = -2 \quad \Rightarrow \quad P(p_s \leq .16) = .0228
\]

Notice the range asked for is equal to \(\pm 2\sigma\) so we already knew the answer = 95.44%.

c) \(P(.14 \leq p_s \leq .26)

\[
z = \frac{.26 - .2}{.02} = 3 \quad \Rightarrow \quad P(p_s \leq .26) = .99865
\]

\[
z = \frac{.14 - .2}{.02} = -3 \quad \Rightarrow \quad P(p_s \leq .14) = .00135
\]

Notice the range asked for is equal to \(\pm 3\sigma\) so we already knew the answer = 99.73%.

Also notice the Z values of these ranges are the multiples 1, 2, and 3. If you see these Z values you know your dealing with \(\sigma\) multiples.
Example:

\[ n = 200 \quad p = .4 \quad P \left( p_s < .43 \right) = ? \]

\[
P \left( p_s < .43 \right) = P \left( \frac{p_s - \mu_{p_s}}{\sigma_{p_s}} < \frac{.43 - .4}{\sqrt{\frac{4(1-.4)}{200}}} \right) = P \left( Z < .87 \right) \approx .8078
\]

Sampling Distribution

\[ \sigma_{p_s} \]

\[ \mu_{p_s} \]

\[ p_s \]

Standardized Normal Distribution

\[ \sigma_{Z} = 1 \]

\[ 0 \quad .87 \]

\[ Z \]

知情的sigma / 百分比规则:

\[ 68\%, \quad 2\% \quad 95\%, \quad 3\% \quad 99.7\% \]

\[ np = (200)(.4) = 80 \]

\[ n(1 - p) = (200)(.6) = 120 \]

\[ p_s = N \left( p, \sqrt{\frac{p(1-p)}{n}} \right) = N \left( .4, \sqrt{\frac{.4 \times .6}{200}} \right) = N \left( .4, .035 \right) \]

Exam

Know the sigma / percentage rules:

\[ \pm \sigma = 68\%, \quad \pm 2\sigma = 95\%, \quad \pm 3\sigma = 99.7\% \]
EXAMPLE CONTINUED

- If the true proportion of voters who support Proposition A is $p = .4$, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?

- i.e.:  
  \[ P(0.40 \leq \hat{p} \leq 0.45) \]

- if $p = .4$ and $n = 200$, what is $P(0.40 \leq \hat{p} \leq 0.45)$?

  Find $\sigma_{\hat{p}}$:
  \[
  \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464
  \]

  Convert to standard normal:
  \[
  P(0.40 \leq \hat{p} \leq 0.45) = P\left(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}\right)
  \]
  \[
  = P(0 \leq Z \leq 1.44)
  \]

  Use standard normal table:  
  \[
  P(0 \leq Z \leq 1.44) = 0.4251
  \]
Sampling from non-Normally Distributed Populations:

**Central Limit Theorem**: as the sample size gets large enough, the sampling distribution of the mean can be approximated by the normal distribution.

Because of the Unbiased property, the mean of any sampling distribution is always equal to the population.

If unbiased, the mean of each sampling distribution is equal to the mean of the population. The variability decreases as the sample size increases.