Applied Statistics

Estimating unknown quantities from a sample

Cesar O. Aguilar
SUNY Geneseo

Portions of these notes were created from *Learning statistics with R* by Danielle Navarro, *Learning statistics with jamovi* by David Foxcroft, and *Introduction to Statistical Thinking* by Benjamin Yakir.

Introduction

- Now that we have some knowledge in probability theory, we can begin to think about the problem of statistical inference
- The main ideas that lie at the heart of inferential statistics are traditionally divided into two "big ideas":
  - estimation
  - hypothesis testing
- The goal in this chapter is to introduce the first of these big ideas, estimation theory
- To that end, we need to first discuss sampling theory because estimation theory doesn't make sense until we understand sampling

Samples, populations, and sampling

- Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely
- In order to talk about "making inferences", we need to be a bit more explicit about what it is that we're drawing inferences from (the sample) and what it is that we're drawing inferences about (the population)
- In almost every situation of interest what we have available to us as researchers is a **sample** of data
- The data set available to us is finite and incomplete since we can't possibly observe the entire population of interest
- A sample was the only thing we were interested in when we covered descriptive statistics
- Our only goal then was to find ways of describing, summarising and graphing that sample

Defining a population

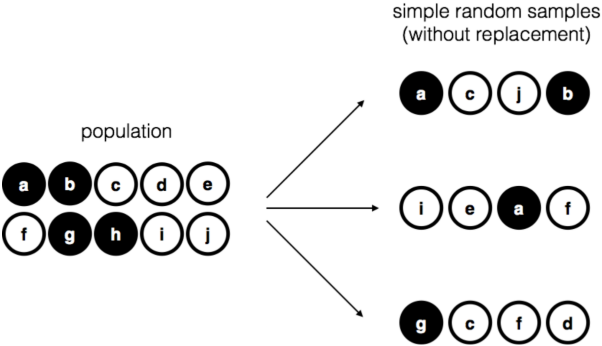
- A **population** is a set of similar items or events which is of interest for some question or experiment
- The items can be a group of existing objects, for example, all registered voters in a particular state, all existing stars in the Milky Way, or all bears in Yellowstone National Park
- The items can also be hypothetical objects or events, for example, all possible outcomes of a coin toss or the set of all possible hands of a poker game
- A population can be a an abstract idea or a concrete collection of objects, but regardless, it refers to all possible items/events of interest

Defining a population

- In many cases, it is not always perfectly clear what the population is
- Suppose we run an experiment using 100 undergraduate students
- The goal of the study is to learn something about human behaviour
- Possible populations include:
  - All undergraduate students at SUNY Geneseo?
  - Undergraduate liberal-arts students in general, anywhere in the world?
  - Americans currently living?
  - Americans of similar ages to the sample?
  - Anyone currently alive?
  - Any human being, past, present or future?
  - Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
  - Any intelligent being?

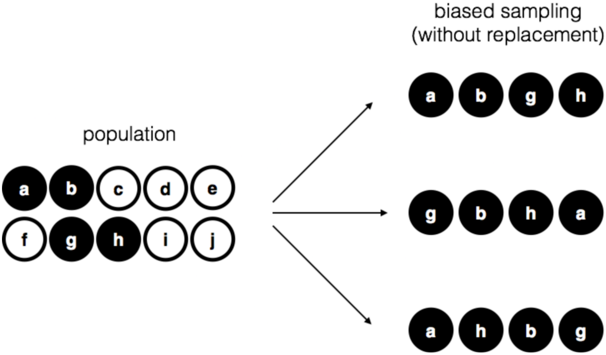
Simple random samples

- The procedure by which a sample is selected from a population is referred to as a **sampling method**
- A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**
- Simple random samples **without replacement**:



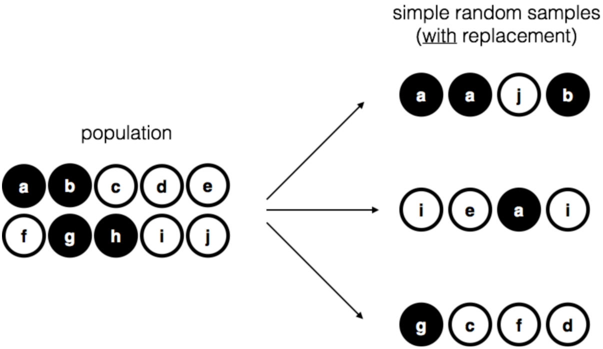
Simple random samples

- **Biased** sampling occurs when some members of a population are systematically more likely to be selected in a sample than others:



Simple random samples

- In simple random samples **with replacement**, each member of the population has the same chance of being selected, but once a member is selected, it is returned to the population and can be selected again:



Simple random samples

- Most statistical theory is based on the assumption that the data arise from a simple random sample with replacement
- In real life this very rarely matters if the population is large
- In this case, the difference between sampling with- and without-replacement is too small to be concerned with

Simple random samples

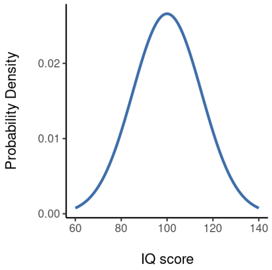
- In some cases, it is impossible to obtain a simple random sample
- Many other sampling methods are available
- **Stratified sampling** - divide the population into groups (or strata) and then randomly select a sample from each group; could lead to **oversampling** because it may deliberately attempt to overrepresent rare groups
- **Snowball sampling** - start with a small sample and then use that sample to recruit more members of the population, very common in social science research; a disadvantage is that the procedure can be unethical if not handled well
- **Convenience sampling** - select a sample that is easy to obtain, for example, by asking people to volunteer; generally non-random and can lead to **bias** in the sample

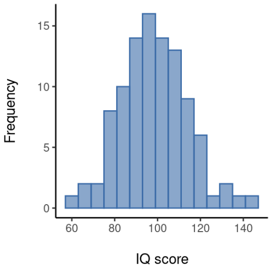
Population parameters and sample statistics

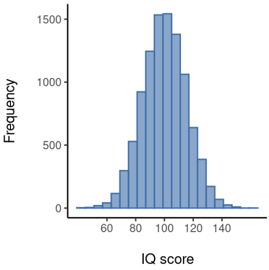
- So what is a population to a statistician?
- To a statistician, which is what we are for the purposes of this course, a population is represented by a probability distribution
- For example, suppose that we are interested in the IQ scores of all students at a particular university
- IQ tests are designed to produce scores that are normally distributed with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$
- These values (μ and σ) are called the **population parameters**
- We then say that $\mu = 100$ is the **population mean** and $\sigma = 15$ is the **population standard deviation**
- As far as we are concerned, the population is completely defined by these two parameters

Population parameters and sample statistics

- If we were to obtain a sample of 100 IQ scores from this population and calculate the mean and standard deviation of the sample, we would obtain values that are different from the population parameters
- For our hypothetical sample of 100 scores, we might obtain a sample mean of $\bar{X} = 98.5$ and a sample standard deviation of $s = 15.9$
- These values are called **sample statistics**

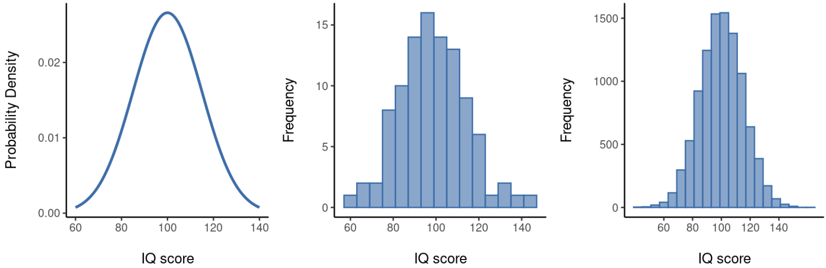






The Law of Large Numbers

- In the IQ example, the sample mean was $\bar{X} = 98.5$
- This is close to the population mean of $\mu = 100$ but what if we wanted a sample mean that was closer to the population mean?
- The answer is that we need to increase the sample size
- Running a simulated experiment of sampling $N = 10,000$ scores from the population and calculating the sample mean and std, we obtain $\bar{X} = 99.65$ and $\sigma = 14.90$



The Law of Large Numbers

- It is intuitively clear that large samples generally give you better information about the population
- This intuition that we all share turns out to be correct, and statisticians refer to it as the **Law of Large Numbers**
- The law of large numbers is a mathematical law that applies to many different sample statistics (not just the mean) but the simplest way to think about it is as a general law about averages
- When applied to the sample mean, the law of large numbers states that as the sample size N tends to infinity ($N \to \infty$), the sample mean $\bar{X}$ will tend to the population mean μ ($\bar{X} \to \mu$)

Sampling distributions

- The law of large numbers is a "long-run guarantee" that in practice is not all that useful because we rarely have access to large samples
- In real life, it would be more useful to be able to learn the behaviour of a sample statistic when it is calculated from a modest sized sample
- This is where **sampling distributions** come in
- Suppose instead that we can only measure the IQ scores of $N = 5$ individuals and we obtain the numbers

$$90, 82, 94, 99, 110$$

- The sample mean is $\bar{X} = (90 + 82 + 94 + 99 + 110)/5 = 95$
- Not surpisingly, this is much less accurate than the sample mean with $N = 10,000$ or even $N = 100$
- Now suppose that we replicate this experiment and measure the IQ scores of another $N = 5$ individuals, and then repeatedly

Sampling distributions

- After performing this experiment 10 times we obtain the following data:

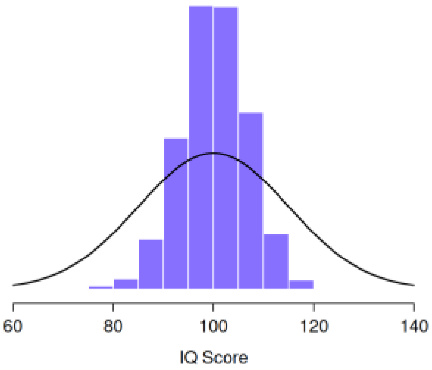
	1	2	3	4	5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

- If we replicated this experiment even further, we would obtain the data points of sample means:

$$95.0, 101.0, 101.6, 103.8, 104.4, 92.4, 106.4, 102.8, 100.4, 105.8, \ldots$$

Sampling distributions

- Here is a histogram of the sample means from the replications:



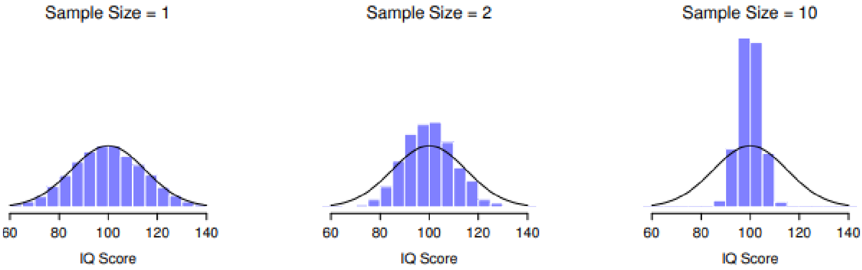
- The average of 5 IQ scores is usually between 80 and 120

Sampling distributions

- This experiment is an example of a **sampling distribution**
- Specifically, this distribution is called the **sampling distribution of the mean**
- Sampling distributions exist for any sample statistic and not just the mean
- For example, the sampling distribution of the median is the distribution of medians obtained from all possible samples of certain size N
- Or, the sampling distribution of the maximum is the distribution of maximums obtained from all possible samples of certain size N
- We do not intend to actually perform repeated experiments and build a histogram of the sample means
- Instead, we will use the theoretical properties of sampling distributions to learn about the behaviour of the sample statistic

Sampling distributions

- The sampling distribution of the mean as the sample size varies:



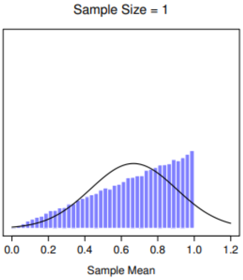
(a) (b) (c)

- As the sampling size increases, the distribution of sample means tend to be fairly tightly clustered around the true population mean

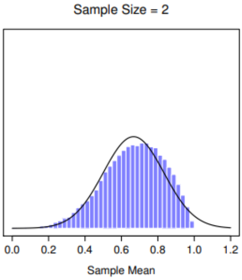
The Central Limit Theorem

- As the sample size increases, the standard deviation of the sampling distribution of the mean decreases
- The standard deviation of the sampling distribution is referred to as the **standard error**, often denoted as SE
- The standard error of the sample mean is often denoted by SEM
- Notice that the sampling distribution of the mean looks very much like a normal distribution
- This is not surprising because IQ scores are normally distributed
- What if the population is not normally distributed?
- The remarkable thing is that no matter what shape your population distribution is, as N increases, the sampling distribution of the mean starts to look more like a normal distribution!

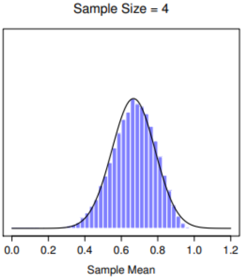
The Central Limit Theorem

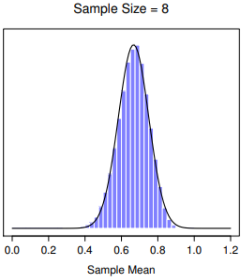


(a)



(b)





The Central Limit Theorem

- It seems like we have evidence for all of the following claims about the sampling distribution of the mean:
  - The mean of the sampling distribution is the same as the mean of the population
  - The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
  - The shape of the sampling distribution becomes normal as the sample size increases

- These facts are summarized in what is called the **Central Limit Theorem** (CLT)

- The CLT tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean is

$$SEM = \frac{\sigma}{\sqrt{N}}$$

The Central Limit Theorem

- From the formula for the standard error of the mean
$$SEM = \frac{\sigma}{\sqrt{N}}$$
we observe that as N increases, the standard error of the mean decreases

- The CLT also tells us that the shape of the sampling distribution becomes normal

- This is one reason why the normal distribution is so important in statistics

- Many experiments measure a characteristic that is a sort of average value of a population and so the distribution of many random variables are approximately normal distributed

- For example, IQ scores can be thought of as an average of many different abilities and so IQ scores are approximately normally distributed

The Central Limit Theorem

- Links to online resources on the Central Limit Theorem:
  - Webassign: The Sampling Distribution
  - Skew the Script: Sampling Distribution for a Mean

Estimating population paramters

- Suppose that we are interested in estimating the mean IQ μ of a certain population
- 100 individuals from the population are randomly selected and their IQ scores are measured; the mean IQ of the **sample** is $\bar{X} = 98.5$
- It is sensible to use the sample mean $\bar{X}$ as an estimate of the population mean μ
- Our estimate for the mean μ is then $\hat{\mu} = 98.5$
- The hat notation is used to indicate that the quantity is an estimate

Symbol	What is it?	Do we know what it is?
$\bar{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of μ	Yes, identical to the sample mean

Estimating the population standard deviation

- If σ is the population standard deviation, we use $\hat{\sigma}$ to denote an estimate for σ
- Following what we did to estimate μ, it seems reasonable to use the sample standard deviation s as an estimate of σ
- Recall that the variance s^2 is given by

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

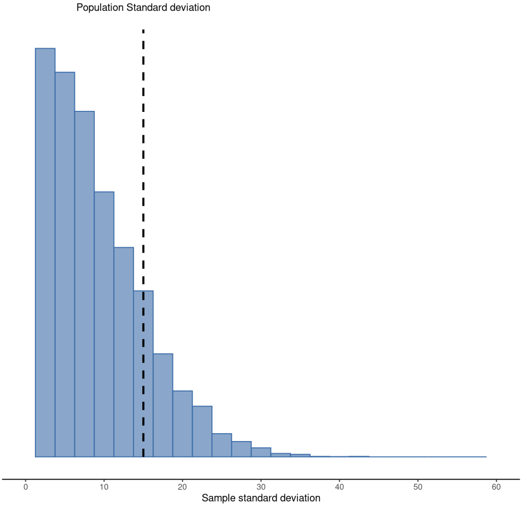
- It turns out that using s as an estimate of σ is not a good idea; in fact s is a **biased** estimator of σ
- Generally, if $\hat{\theta}$ is an estimator of θ, then $\hat{\theta}$ is **unbiased** if the sampling distribution of $\hat{\theta}$ is centered at θ
- The sampling distribution of $\bar{X}$ is centered at μ and so $\bar{X}$ is an unbiased estimator of μ

Estimating the population standard deviation

- It turns out that the sampling distribution of the variance s^2 is centered at
$$\frac{N-1}{N}\sigma^2$$
- Thus, the sample variance s^2 is a **biased** estimator of σ^2
- On the other hand, the sampling distribution of the estimator
$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$
is centered at σ^2
- Therefore, an improved estimator of σ is
$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$
- Many people refer to $\hat{\sigma}$ as the **sample standard deviation**

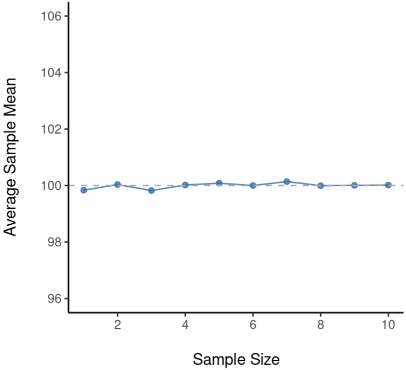
Estimating the population standard deviation

Sampling distribution of s for $N = 2$ centered at 8.5; $\sigma = 15$

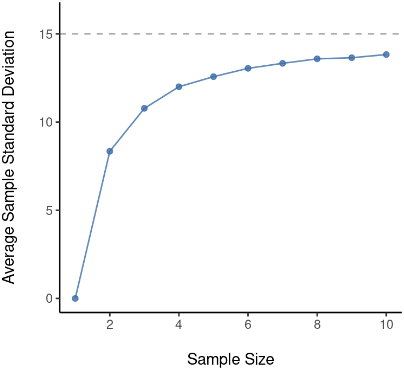


Estimating the population standard deviation

- (a) The sample mean is an unbiased estimator of the population mean; (b) The sample standard deviation is a biased estimator of the population standard deviation



(a)



(b)

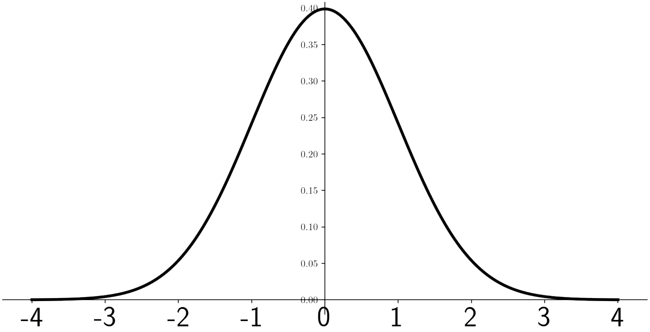
Estimating population parameters

Symbol	What is it?	Do we know what it is?
s	Sample standard deviation	Yes, calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes, but not the same as the sample standard deviation

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

The standard normal distribution $N(0,1)$

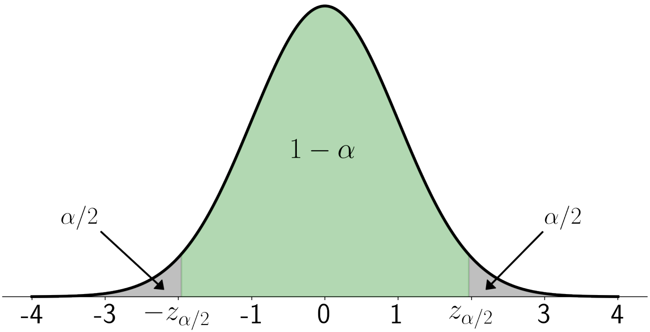
- We now describe how to quantify the amount of uncertainty that attaches to our estimates
- For this we use the CTL and the **standard normal distribution**
- The standard normal distribution is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$



The standard normal distribution $N(0,1)$

- There is a $100(1-\alpha)\%$ probability that Z lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



- For $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \Longrightarrow P(-1.96 \leq Z \leq 1.96) = 0.95$

Estimating a confidence interval for μ

- Let's return to describing how to quantify the amount of uncertainty that attaches to our estimates
- For example, it would be nice to be able to say that there is a 95% chance that the true mean μ lies between 109 and 121
- The name for this is a **confidence interval** for the mean
- Suppose that the true mean and standard deviation of a population are μ and σ
- If we gather a random sample of size $N \geq 30$, the sampling distribution of $\bar{X}$ is approximately normal with mean μ and standard deviation $SEM = \sigma/\sqrt{N}$
- Therefore, the random variable

$$Z = \frac{\bar{X} - \mu}{SEM}$$

has (approximately) a standard normal distribution: $Z \sim N(0, 1)$

Estimating a confidence interval for μ

- Therefore, there is a $100(1-\alpha)\%$ probability that Z lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$:

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$$

- After some algebraic manipulations, we get:

$$-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{SEM} \leq z_{\alpha/2}$$

- After some algebraic manipulations, we get:

$$\bar{X} - (z_{\alpha/2} \cdot SEM) \leq \mu \leq \bar{X} + (z_{\alpha/2} \cdot SEM)$$

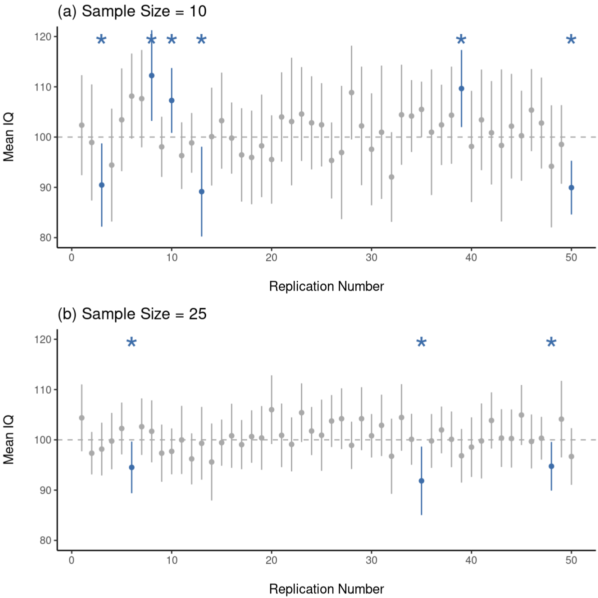
- The interval $\bar{X} \pm (z_{\alpha/2} \cdot SEM)$ is called a $100(1-\alpha)\%$ confidence interval for μ
- The term $E = z_{\alpha/2} \cdot SEM$ is called the **margin of error** or just **error**
- As an example, the 95% confidence interval is

$$\bar{X} \pm (1.96 \cdot SEM)$$

Interpreting a confidence interval

- We need to be careful how we interpret a confidence interval
- A confidence interval is not a prediction; we are not saying that the true mean μ lies in $\bar{X} \pm (1.96 \cdot SEM)$ with 95% probability
- This interpretation is not consistent with the frequentist interpretation of probability
- Instead, we are saying that if we were to gather many random samples of size N from the population, and construct a 95% confidence interval for μ in each case, then 95% of the intervals would contain μ
- To say that **we are 95% confident that the unknown μ lies in the interval $\bar{X} \pm (1.96 \cdot SEM)$** is to say that "We got these numbers using a method that gives correct results 95% of the time"
- We don't know whether the 95% confidence interval from a particular sample is one of the 95% that capture μ or one of the unlucky 5% that miss

Estimating a confidence interval for μ



Estimating confidence intervals in general

- The procedure described above can be used to construct confidence intervals for any population parameter θ and not just μ
- One main assumption in our procedure was that the sampling distribution for the point estimate be reasonably modeled as normal
- Then, a $100(1-\alpha)\%$ confidence interval for the unknown parameter θ can be constructed as

$$\hat{\theta} \pm (z_{\alpha/2} \cdot SE(\hat{\theta}))$$

  where $SE(\hat{\theta})$ is the standard error of the point estimate $\hat{\theta}$
- A 95% confidence interval for θ is $\hat{\theta} \pm (1.96 \cdot SE(\hat{\theta}))$
- A 99% confidence interval for θ is $\hat{\theta} \pm (2.58 \cdot SE(\hat{\theta}))$
- $(1-\alpha)$ is called the **confidence level**, $z_{\alpha/2}$ is called the **critical value**, and the **margin of error** $z_{\alpha/2} \cdot SE(\hat{\theta})$

Example: Confidence interval for μ

Example. A simple random sample (SRS) of $n = 74$ observations produced a sample mean of $\bar{x} = 110.73$ from a population known to have a standard deviation of $\sigma = 11$. Find a 95% confidence interval for the population mean μ.

- The z-value for a 95% confidence interval is $z = 1.96$
- The standard error is $SE = \sigma/\sqrt{n} = 11/\sqrt{74} = 1.27872$
- The error is $E = z \cdot SE = 1.96 \cdot 1.27872 = 2.506299$
- The confidence interval is then

$$\bar{x} \pm E = 110.73 \pm 2.506299 = (108.22, 113.24)$$

- We are 95% confident that the true mean μ lies in the interval $(108.22, 113.24)$

Example: Confidence interval for μ

Example. A simple random sample (SRS) of $n = 64$ observations produced a sample mean of $\bar{x} = 33$ from a population known to have a variance of $\sigma^2 = 256$. Find a 90% confidence interval for the population mean μ.

- The z-value for a 90% confidence interval is $z = 1.645$
- The standard error is $SE = \sigma/\sqrt{n} = \sqrt{256}/\sqrt{64} = 2$
- The error is $E = z \cdot SE = 1.645 \cdot 2 = 3.29$
- The confidence interval is then

$$\bar{x} \pm E = 33 \pm 3.29 = (29.71, 36.29)$$

- Our estimate for the mean is $\hat{\mu} = 33$ and we are 90% confident that the true mean μ lies in the interval $(29.71, 36.29)$

Sample size for a confidence interval

- From the margin of error formula

$$E = z \cdot \frac{\sigma}{\sqrt{n}}$$

  we can solve for n to get

$$n = \left[\frac{z \cdot \sigma}{E}\right]^2$$

- This can be used to determine the sample size needed to achieve a desired margin of error E

Example: Sample size for a confidence interval

Example. The registrar's office wants to estimate the average amount of time it takes students to walk from one class to the next. They want to be 95% confident that the true average is within 0.3 minutes of the sample mean. The standard deviation of the population is 1.5 minutes. How many students should be surveyed?

- We are given that $\sigma = 1.5$ and $E = 0.3$
- The critical value for a 95% confidence interval is $z = 1.96$
- The sample size needed is then

$$\begin{aligned} n &= \left[\frac{z \cdot \sigma}{E}\right]^2 \\ &= \left[\frac{1.96 \cdot 1.5}{0.3}\right]^2 \\ &= 96.04 \end{aligned}$$

- We round up to $n = 97$

Binomial Distribution Revisited

- Suppose that the probability of success of a single experiment (or trial) is θ
- Then if X denotes the number of successes in n trials, then X has a binomial distribution with parameters n and θ:

$$X \sim B(n, \theta)$$

- Problems involving **proportions** are often modeled using the binomial distribution
- For example, suppose that we want to estimate the **proportion** of people who like a particular brand of soft drink
- We could randomly select a sample of n people and ask them if they like the soft drink
- If X denotes the number of people who like the soft drink, then X has a binomial distribution with parameters n and θ

Binomial Distribution Revisited

- When dealing with proportions, it is often convenient to use the notation p instead of θ to denote the probability of success in a single trial
- The mean of a binomial distribution $B(n, p)$ is

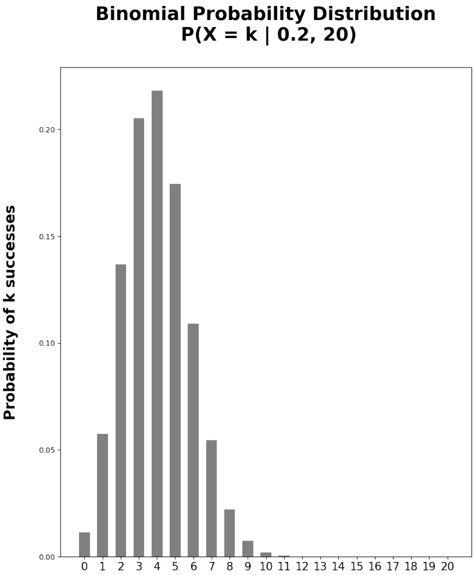
$$\mu = np$$

- And the standard deviation is

$$\sigma = \sqrt{np(1-p)}$$

- We often write $q = (1-p)$ and then $\sigma = \sqrt{npq}$
- q is then the probability of failure in a single trial

The binomial distribution: $\theta = 0.2$, $N = 20$



Proportions

- A random variable X that has a binomial distribution $B(n, p)$ can be thought of as the sum of n independent random variables $X_1, X_2, \ldots, X_n$
- Each random variable X_i is the result of a single trial and has probability distribution:

Event	Probability
1 (success)	$P(X_i = 1) = p$
0 (failure)	$P(X_i = 0) = 1 - p$

- Thus $X = X_1 + X_2 + \cdots + X_n$
- Let $Y = \frac{1}{n}X$ be the proportion of successes in n trials
- Then Y is a random variable with parameters $\mu = p$ and

$$\sigma = \sqrt{pq} = \sqrt{p(1-p)}$$

Proportions

- By the Central Limit Theorem, Y can be approximated by a normal distribution with mean $\mu = p$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

- In this case, the standard error is called the **standard error of the proportion**:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- We can now proceed as before to find confidence intervals for the population mean $\mu = p$
- The estimator for p is the sample proportion $\hat{p} = \frac{X}{n}$
- The $100(1-\alpha)\%$ confidence interval for p is then

$$\hat{p} \pm z_{\alpha/2} \cdot SE = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: Confidence interval for a proportion

Example. A random sample of 400 students is selected. Of these students, 136 like the new cafeteria food. Construct a 95% confidence interval for the proportion of students who like the new cafeteria food.

- We are given that $n = 400$ and $X = 136$
- The sample proportion is $\hat{p} = \frac{X}{n} = 0.34$
- The critical value for a 95% confidence interval is $z = 1.96$
- The error is

$$E = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{0.34(1-0.34)}{400}} = 0.046423$$

- The 95% confidence interval is then

$$\hat{p} \pm E = 0.34 \pm 0.046423 = (0.294, 0.386)$$

Sample size for a Proportion

- From the margin of error formula

$$E = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  we can solve for n to find the sample size needed to achieve a given margin of error E

- Solving for n gives

$$n = \hat{p}(1-\hat{p}) \left[\frac{z}{E}\right]^2$$

- To use this formula we need a value for $\hat{p}$ which involves obtaining a sample!

- There are two ways to proceed:
  - If we know a range of values of the true proportion p then we choose p closest to 0.5
  - If p is completely unknown, then we choose $p = 0.5$

Example: Sample size for a Proportion

Example. A survey is to be conducted to determine the proportion of students who like the new cafeteria food. It is desired to estimate the proportion with a 95% confidence level. The margin of error is to be no more than 0.04. How large a sample is needed if

- p is known to be between 0.1 and 0.25;
- p is completely unknown?

- We are given that $E = 0.04$ and $z = 1.96$
- If p is known to be between 0.1 and 0.25, then we choose $p = 0.25$ and compute

$$n = 0.25(1 - 0.25)\left[\frac{1.96}{0.04}\right]^2 \approx 451$$

- If p is completely unknown, then we choose $p = 0.5$ and compute

$$n = 0.5(1 - 0.5)\left[\frac{1.96}{0.04}\right]^2 \approx 601$$