Confidence Intervals

553.211 Intro to Prob. & Stats.

Johns Hopkins University

EN.553.211

$V(\overline{X})$ for finite population

- Recall that when the population size N is much larger than the sample size n, it allows us to say that the sample points $X_1, X_2, \ldots, X_n$ are approximately independent even though the sampling is performed without replacement.

$V(\overline{X})$ for finite population

- Recall that when the population size N is much larger than the sample size n, it allows us to say that the sample points $X_1, X_2, \ldots, X_n$ are approximately independent even though the sampling is performed without replacement.
- The independent gives us

$$V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$
$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$V(\overline{X})$ for finite population

- Recall that when the population size N is much larger than the sample size n, it allows us to say that the sample points $X_1, X_2, \ldots, X_n$ are approximately independent even though the sampling is performed without replacement.
- The independent gives us

$$V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$
$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- A more careful analysis is needed for small N. In this case $X_1, X_2, \ldots, X_n$ are not independent (due to the non-replacement). It leads to the following, more accurate formula

$$V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

$V(\overline{X})$ for finite population

Fact

If a random sample of size n are selected from a finite population of size N with mean μ and std. dev. σ, then the sampling distribution of $\overline{X}$ will have

$$E(\overline{X}) = \mu \quad \text{and} \quad V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

where the factor $\frac{N-n}{N-1}$ is called **finite population correction**

Notice $E(\overline{X}) = \mu$ doesn't need the independence of $X_1, X_2, \ldots, X_n$

$V(\overline{X})$ for finite population

Example (Sampling from a Small Population)

A population consists of $N = 5$ numbers $\{3, 6, 9, 12, 15\}$. If a random sample X_1, X_2, X_3 $(n = 3)$ is selected without replacement, find the distribution of $\overline{X}$ and calculate $\mu_{\overline{X}}$ and $\sigma^2_{\overline{X}}$ (mean and variance of the sample average $\overline{X}$). Compare these values with the population mean and variance μ and σ^2.

$V(\overline{X})$ for finite population

There are $\binom{5}{3} = 10$ equally likely ways we could pick 3 numbers out of 5:

Sample #	X_1, X_2, X_3
1	3,6,9
2	3,6,12
3	3,6,15
4	3,9,12
5	3,9,15
6	3,12,15
7	6,9,12
8	6,12,15
9	6,9,15
10	9,12,15

$V(\overline{X})$ for finite population

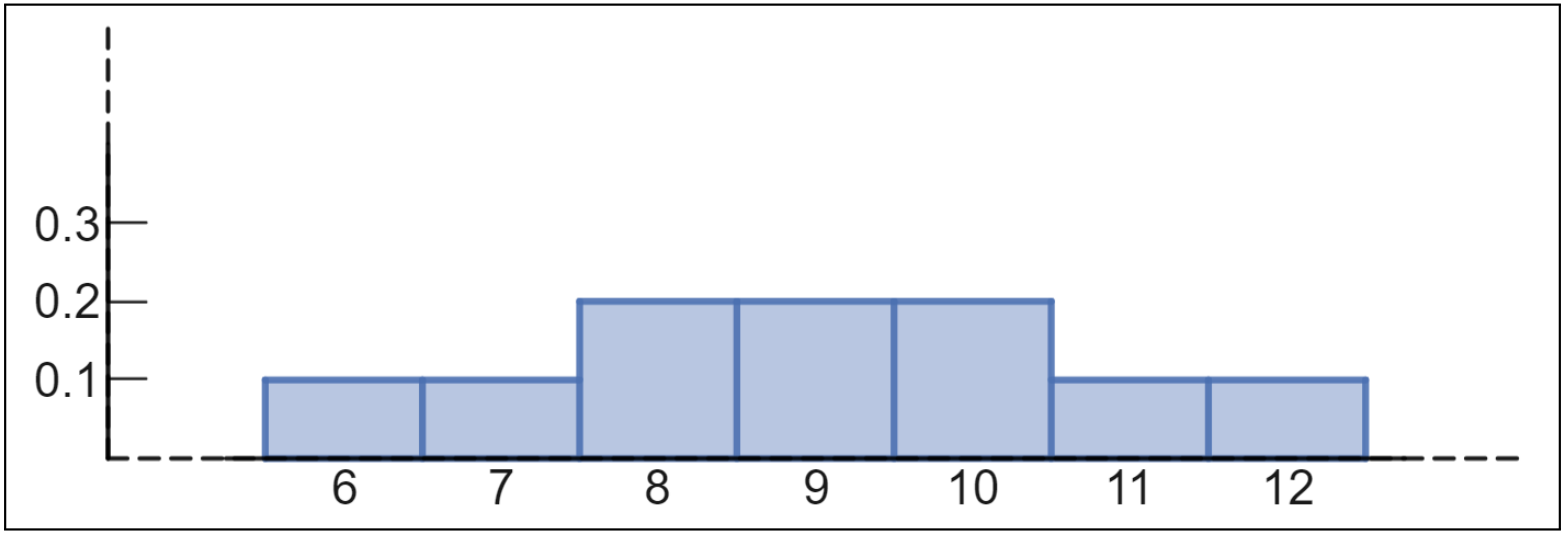
For each of these sample we compute the mean $\overline{X}$:

Sample #	X_1, X_2, X_3	$\overline{X}$
1	3,6,9	6
2	3,6,12	7
3	3,6,15	8
4	3,9,12	8
5	3,9,15	9
6	3,12,15	10
7	6,9,12	9
8	6,12,15	10
9	6,9,15	11
10	9,12,15	12

$V(\overline{X})$ for finite population

Here we find the pmf and probability histogram of $\overline{X}$:

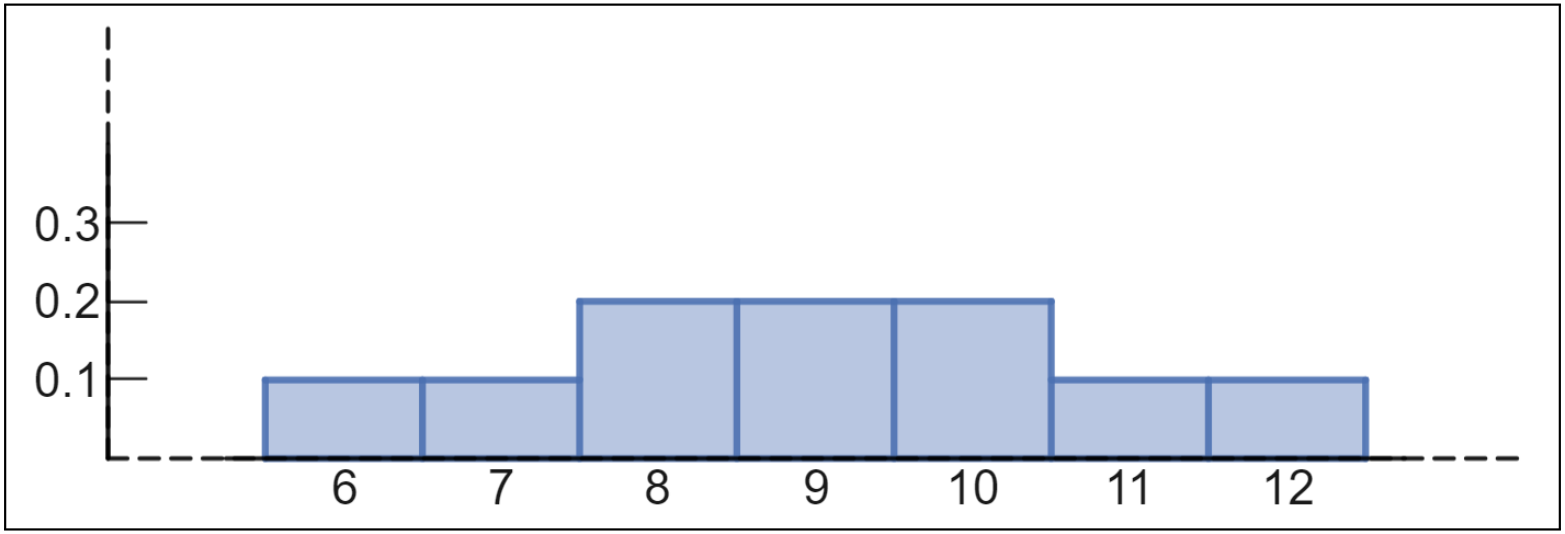
$\overline{x}$	6	7	8	9	10	11	12
$f(\overline{x})$	0.1	0.1	0.2	0.2	0.2	0.1	0.1



$V(\overline{X})$ for finite population

Here we find the pmf and probability histogram of $\overline{X}$:

$\overline{x}$	6	7	8	9	10	11	12
$f(\overline{x})$	0.1	0.1	0.2	0.2	0.2	0.1	0.1



Next, we compute

$$\mu_{\overline{X}} = 0.1(6 + 7 + 11 + 12) + 0.2(8 + 9 + 10) = 9$$

$V(\overline{X})$ for finite population

Let's compare this with the the population mean:

$$\mu = \frac{1}{5}(3 + 6 + 9 + 12 + 15) = \frac{45}{5} = 9.$$

So $\mu_{\overline{X}} = \mu$.

$V(\overline{X})$ for finite population

We now compute the variance of $\overline{X}$:

$$\begin{aligned}(\sigma_{\overline{X}})^2 &= \sum_{\text{all } \overline{x}} (\overline{x} - 9)^2 f(\overline{x}) \\ &= 0.1((-3)^2 + (-2)^2 + 2^2 + 3^2) + 0.2(1^2 + 0^2 + 1^2) \\ &= 3.\end{aligned}$$

$V(\overline{X})$ for finite population

We now compute the variance of $\overline{X}$:

$$\begin{aligned}(\sigma_{\overline{X}})^2 &= \sum_{\text{all } \overline{x}} (\overline{x} - 9)^2 f(\overline{x}) \\ &= 0.1((-3)^2 + (-2)^2 + 2^2 + 3^2) + 0.2(1^2 + 0^2 + 1^2) \\ &= 3.\end{aligned}$$

We then compute the variance of the population

$$\begin{aligned}\sigma^2 &= \frac{1}{5}\sum_{i=1}^{5}(x_i - 9)^2 \\ &= \frac{1}{5}((-6)^2 + (-3)^2 + 0^2 + 3^2 + 6^2) \\ &= 18.\end{aligned}$$

$V(\overline{X})$ for finite population

We now compute the variance of $\overline{X}$:

$$\begin{aligned}(\sigma_{\overline{X}})^2 &= \sum_{\text{all } \overline{x}} (\overline{x} - 9)^2 f(\overline{x}) \\ &= 0.1((-3)^2 + (-2)^2 + 2^2 + 3^2) + 0.2(1^2 + 0^2 + 1^2) \\ &= 3.\end{aligned}$$

We then compute the variance of the population

$$\begin{aligned}\sigma^2 &= \frac{1}{5}\sum_{i=1}^{5}(x_i - 9)^2 \\ &= \frac{1}{5}((-6)^2 + (-3)^2 + 0^2 + 3^2 + 6^2) \\ &= 18.\end{aligned}$$

With $n = 3$ and $N = 5$ we have

$$\sigma^2_{\overline{X}} = 3 = \frac{18}{3} \cdot \frac{2}{4} = \frac{\sigma^2}{3} \cdot \frac{5-3}{5-1}$$

$V(\overline{X})$ for finite population

Remark

Note that for $N \gg n$ we have $\frac{N-n}{N-1} \approx 1$, then

$$V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \approx \frac{\sigma^2}{n}$$

It fits our result for large N. And it means in this case the variance of the sample mean does not depend on the population size N.

Confidence interval

- Suppose we want to know a parameter of population.
- We can get a sample of the population, and use the corresponding statistic as the point estimator of the parameter.
- Actually, for a given sample, we can also give an interval to estimate the parameter:
  Confidence intervals: it provides a margin of error (i.e. a measure of the precision) of our estimate.

Confidence interval

Suppose $X \sim N(\mu, \sigma)$. For what constant c do we have

$$P(\mu - c \cdot \sigma \leq X \leq \mu + c \cdot \sigma) = 95\%?$$

Confidence interval

Suppose $X \sim N(\mu, \sigma)$. For what constant c do we have

$$P(\mu - c \cdot \sigma \leq X \leq \mu + c \cdot \sigma) = 95\%?$$

We have

$$\begin{aligned} 0.95 &= P(\mu - c \cdot \sigma \leq X \leq \mu + c \cdot \sigma) \\ &= P(-c \leq \frac{X - \mu}{\sigma} \leq c) \\ &= P(-c \leq Z \leq c) \end{aligned}$$

Here $Z \sim N(0, 1)$.

Confidence interval

Suppose $X \sim N(\mu, \sigma)$. For what constant c do we have

$$P(\mu - c \cdot \sigma \leq X \leq \mu + c \cdot \sigma) = 95\%?$$

We have

$$\begin{aligned} 0.95 &= P(\mu - c \cdot \sigma \leq X \leq \mu + c \cdot \sigma) \\ &= P(-c \leq \frac{X - \mu}{\sigma} \leq c) \\ &= P(-c \leq Z \leq c) \end{aligned}$$

Here $Z \sim N(0, 1)$.
Now we find the constant c such that $P(-c \leq Z \leq c) = 0.95$.
c should be $z_{\alpha/2}$, where $\alpha = 1 - 0.95 = 0.05$

For $\alpha = 0.05$, $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$, then we have

$$P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) \approx 95\%$$

Confidence interval

More general, we have

Fact

Suppose $X \sim N(\mu, \sigma)$, *then we have*

$$P(\mu - z_{\alpha/2} \times \sigma \leq X \leq \mu + z_{\alpha/2} \times \sigma) = 1 - \alpha$$

For any $\alpha \in [0, 1]$

Confidence interval

Definition

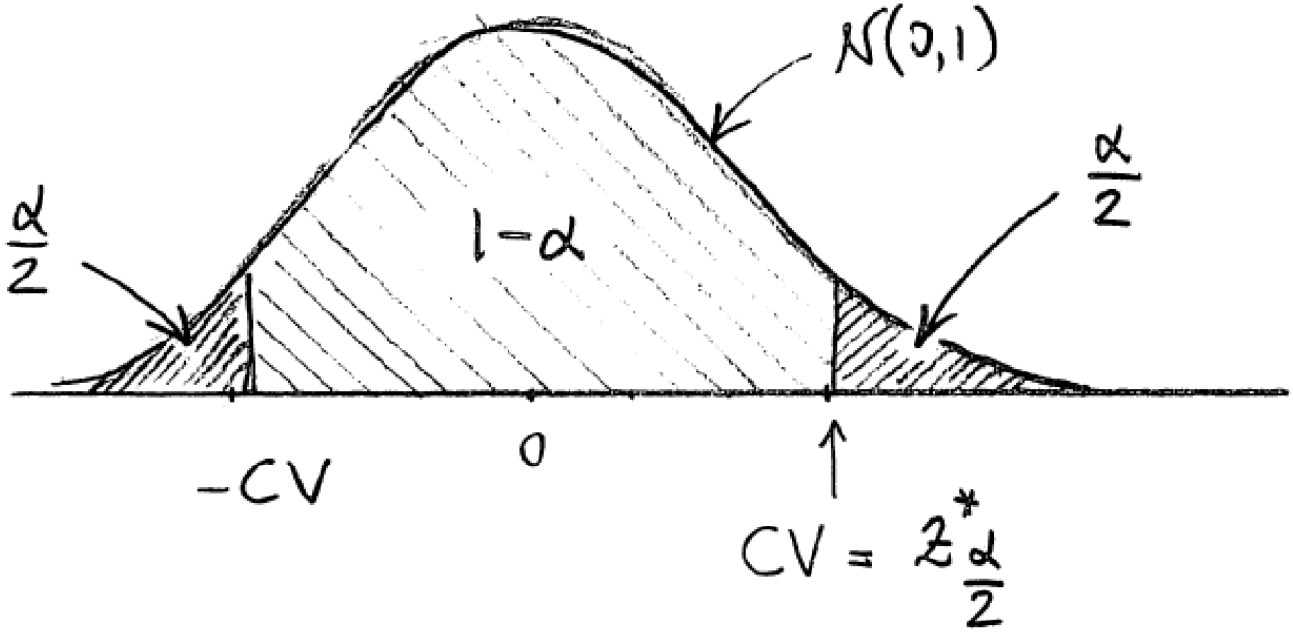
- $1-\alpha$ (area of the central lobe of distribution) is the **confidence**.
- $z_{\alpha/2}$ is called the $(1-\alpha)$-**confidence critical value**. $z_{\alpha/2}$ is the number such that

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

where $Z \sim N(0,1)$

E.g. $\alpha = 0.05, z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$

Confidence interval



Confidence interval

confidence → $1-\alpha$	$\frac{\alpha}{2}$	$z^*_{\frac{\alpha}{2}}$ ← $(1-\alpha)$-confidence critical value
0.9	0.05	1.65
0.95	0.025	1.96
0.99	0.005	2.575
0.995	0.0025	2.81

As $1-\alpha$ increases, the critical value $z_{\alpha/2}$ is "pushed out" and increases too.

Sampling distribution of sample mean

Theorem (Sampling distribution of $\overline{X}$)

We gather a random sample of size n from the population of size N ($n \ll N$). The sampling distribution of the sample mean $\overline{X}$ has the following properties:

1. *Mean:* $E(\overline{X}) = \mu$
2. *Standard deviation (standard error)* $SE(\overline{X}) = SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
3. *Shape:*
   1. *If the population is normal distributed, then*
   $$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
   2. *Of if n are sufficiently large ($n \geq 30$), regardless of the population distributions, by CLT we have*
   $$\overline{X} \overset{approx.}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Confidence interval of population mean

Theorem

Under the above assumptions, we have

$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

E.g. $P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$

Confidence interval of population mean

Definition (Confidence interval of population mean)

Given a population with population mean μ and population std. dev. σ. For a given sample of size n with sample mean $\overline{x}$, the interval

$$CI_{(1-\alpha)\times 100\%} = \left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is called the $(1-\alpha)\times 100\%$ **confidence interval for the population mean** μ.

Confidence interval of population mean

Remark

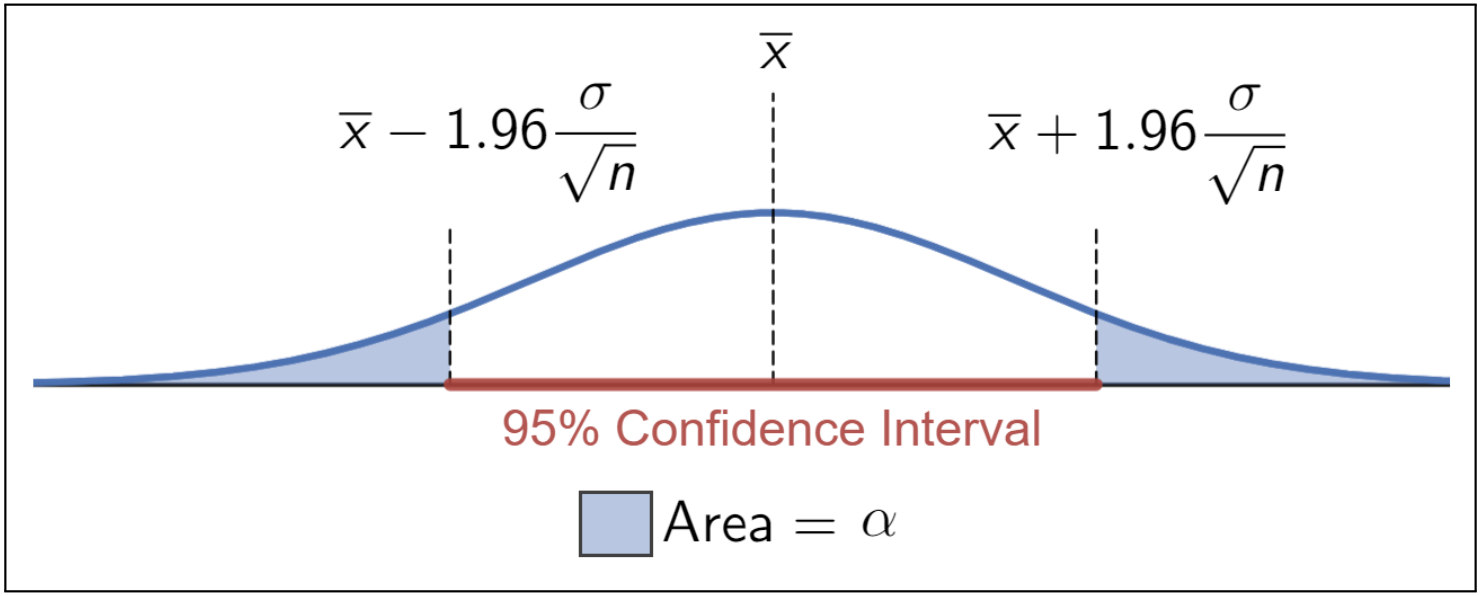
- *It's an interval centered at* $\overline{x}$*, with total width* $2 \times z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- *The theorem says that, with probability* $(1-\alpha)$*, the population mean* μ *belongs to the confidence interval.*
- *In statistical language, we say "With confidence* $(1-\alpha)\times 100\%$*, we have that*

$$\mu \in CI_{(1-\alpha)\times 100\%} = \left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]"$$

E.g. Let's fix $1-\alpha = 0.95$. "With confidence 95%, we have that

$$\mu \in CI_{95\%} = \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]"$$

Confidence interval of population mean



Confidence interval of population mean

Theorem (Confidence interval of μ)

Suppose the population std. dev. σ is known.

- *If the population distribution is normal*
- *Or if the sample size is large enough ($n \geq 30$), then regardless of the population distribution*

We have with $(1-\alpha)\times 100\%$ confidence that

$$\mu \in CI_{(1-\alpha)\times 100\%} = \left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Confidence interval of population mean

- Suppose μ is unknown and we want to give an interval to estimate it. The theorem says that with probability ("confidence") $1-\alpha$ we have that the population mean μ falls within this interval.
- Notice in the above formula of CI we used σ, but in general we may not know the value of σ, so the confidence interval cannot be computed.
  So, we proceed as follows: from the sample $x_1, x_2, \ldots, x_n$ we first compute the sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$, and we use s instead of σ in the confidence interval. Recall that s is a good estimator of σ.

Confidence interval of population mean

- When we have $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$, we have

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

  Thus we get the critical value for the CI from the standard normal distribution.

- If n is large enough (say $n \geq 30$), when we use S instead of σ, the above term will still have the same distribution (still normal), i.e.

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$$

  So we can still get the critical value from the standard normal distribution.

- Notice that when $n < 30$, the distribution changes (not normal).

Confidence interval of population mean

Therefore, the $(1-\alpha)\times 100\%$ confidence interval for μ becomes

Theorem (Confidence interval of μ)

Suppose the population std. dev. σ is unknown.

- *If the sample size is large enough ($n \geq 30$), then regardless of the population distribution*

We have with $(1-\alpha)\times 100\%$ confidence that

$$\mu \in CI_{(1-\alpha)\times 100\%} = \left[\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Confidence Interval

Definition (Confidence Interval)

An **point estimator** is the corresponding statistic of a population parameter (such as sample mean, proportion). A $100 \times (1-\alpha)\%$ **confidence interval** (or **CI** for short) is defined as

$$(\text{Point Estimator}) \pm \underbrace{CV \times (SE \text{ of Point Estimator})}_{\text{this part is called the } \textbf{Margin of Error}}$$

where CV is the $(1-\alpha)$-critical value.

Definition

$(1-\alpha)\%$ **margin of error**, or simple **margin of error** is defined as half the length of $CI_{(1-\alpha)\times 100\%}$

For μ when σ is unknown, $ME = z_{\alpha/2}\frac{s}{\sqrt{n}}$

Confidence Interval

We will learn how to construct confidence intervals for different population parameters (population mean, population proportion . . .). Each one depends on the sampling distribution of the corresponding estimator.

Confidence interval of population mean

- Standard error of $\overline{X}$, as an estimator of μ

$$SE(\overline{X}) = SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

- Estimated standard error of $\overline{X}$ (if σ is unknown)

$$SE(\overline{X}) \approx \frac{s}{\sqrt{n}}$$

- $(1 - \alpha) \times 100\%$ margin of error

$$ME(\overline{X}) = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Estimated $(1 - \alpha) \times 100\%$ margin of error (if σ is unknown)

$$ME(\overline{X}) \approx z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Confidence interval of population mean

Example

A dietitian selected a random sample of $n = 100$ male adults and found that their average daily intake of dairy products was $\overline{x} = 7.56$ grams/day with a standard deviation of $s = 35$ grams/day. Use this sample information to construct a 95% CI for the mean daily intake of dairy products for men.

Confidence interval of population mean

We first check $n = 50 \geq 30$, then we can compute the CI.
$\alpha = 0.05, z_{\alpha/2} = z_{0.025} \approx 1.96$
$ME = z_{\alpha/2}\frac{s}{\sqrt{n}} \approx 1.96\frac{35}{\sqrt{100}} \approx 6.86$
The 95% CI is

$$\overline{x} \pm ME \approx 7.56 \pm 6.86 \approx [0.7, 14.42]$$

We have with 95% confidence that $\mu \in [0.7, 14.42]$

Interpretation of confidence interval

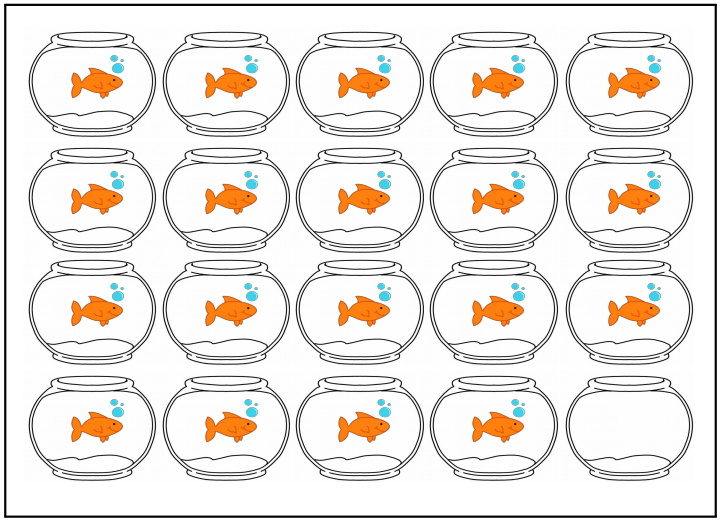
- Confidence intervals, like sample statistics, vary from sample to sample.
- Given a sample, the confidence interval is created to tell us about the location of the population parameter.
- The level of confidence tells us how much we trust our confidence interval (ie, how confident we are that our interval contains the true parameter).
- By convention, many disciplines use 95% confidence intervals. This means that in repeated samples, we would be certain that 95% of the possible samples would produce intervals that contain the true parameter, and 5% of intervals will miss it.

Interpretation of confidence interval

- After we obtain the CI for a given sample, we have no way of knowing if our interval captures the parameter or misses it, but we know that 95% of the time we capture it.
- So we say "with confidence 95%, $CI_{95\%}$ contains the true parameter".
- This does **NOT** mean that there is a 95% chance that the parameter is in our interval. For a particular sample, either the parameter is in the CI or it is not (i.e. For the CI for a given sample, either there is a 100% chance or a 0% chance that the parameter is in the interval).

Interpretation of confidence interval

Helpful metaphor for confidence level - Goldfish:
19 out of 20 bowls have a goldfish (95%)



- Now, I randomly give you a jar, and you look at it (you have an interval, and either the parameter is in it, or it is not).
- If you tell me there is a 95% chance that you are holding a goldfish, that doesn't make sense! Either you're holding a goldfish, or you're not.

Choosing a confidence level

- If I want to create a useful interval, I will always have some *uncertainty* about whether or not the interval contains the true parameter.
- If I want to be really confident that my interval contains the parameter, I will need a pretty wide interval.
- For example, we can always create a 100% confidence interval for a population mean: we are 100% confident that the population mean is in $\mathbb{R}$. But it doesn't help us to understand where is the population mean.

How big is the confidence interval

- Larger level of confidence requires wider confidence intervals.
  Larger level of confidence $\Rightarrow$ larger critical value
- Larger sample sizes yields a more narrow confidence interval.
  Larger sample sizes $\Rightarrow$ smaller SE
  (larger samples give estimates that are less variable)

s^2 is the unbiased estimator of σ^2

- $\overline{X}$ is an unbiased estimator of μ, because $E(\overline{X}) = \mu$
- How about s^2 as the estimator of σ^2?

s^2 is the unbiased estimator of σ^2

$$
\begin{aligned}
\mathrm{E}\left(s^2\right) &= \mathrm{E}\left(\sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n-1}\right) \\
&= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} \mathrm{E}\left(X_i^2\right) - n\mathrm{E}\left(\bar{X}^2\right)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} \left\{\mathrm{V}\left(X_i\right) + \left[\mathrm{E}\left(X_i\right)\right]^2\right\} - n\left\{\mathrm{V}(\bar{X}) + \left[\mathrm{E}(\bar{X})\right]^2\right\}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} \left\{\sigma^2 + \mu^2\right\} - n\left\{\frac{\sigma^2}{n} + \mu^2\right\}\right] \\
&= \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}
$$

s^2 is the unbiased estimator of σ^2

- Thus s^2 is an unbiased estimator of σ^2
- So when we don't have knowledge of the population std. dev. σ, it is reasonable for us to use s instead of σ.

Difference between two population means

Question

What if we want to compare means from two different populations. E.g. Is the avg. MCAT score higher for biology majors or biochemistry majors? Or are they the same? How can we test this?

Difference between two population means

When sampling from two different populations, we will index them by Pop. 1 and Pop. 2. Note that it doesn't matter which population you call 1 & 2, so long as you keep it consistent in the formulas!

Difference between two population means

When sampling from two different populations, we will index them by Pop. 1 and Pop. 2. Note that it doesn't matter which population you call 1 & 2, so long as you keep it consistent in the formulas!

	Pop. 1	**Pop.** 2
Mean	μ_1	μ_2
Std. Dev.	σ_1	σ_2

	Sample 1	**Sample** 2
Mean	$\overline{x}_1$	$\overline{x}_2$
Std. Dev.	s_1	s_2
Sample Size	n_1	n_2

Difference between two population means

- Previously we used $\overline{X}_1$ as an estimator of μ_1, and $\overline{X}_2$ as an estimator of μ_2
- We will use $\overline{X}_1 - \overline{X}_2$ as an estimator of $\mu_1 - \mu_2$. $\overline{X}_1 - \overline{X}_2$ is a random variable!

Sampling distribution of difference between two sample means

Theorem (Sampling distribution of $\overline{X}_1 - \overline{X}_2$)

Given two populations of size N_1 and N_2, we gather two independent random samples from them of size n_1 and n_2, respectively ($n_1 \ll N_1, n_2 \ll N_2$). The sampling distribution of the difference $\overline{X}_1 - \overline{X}_2$ has the following properties:

1. *Mean:* $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$
2. *Standard deviation (standard error):*
$$SE(\overline{X}_1 - \overline{X}_2) = SD(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
3. *Shape:*
    1. *If the two populations are normal distributed, then*
    $$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$
    2. *Of if n_1 and n_2 are sufficiently large ($n_1 \geq 30, n_2 \geq 30$), regardless of the population distributions, by CLT we have*
    $$\overline{X}_1 - \overline{X}_2 \overset{approx.}{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Confidence interval of difference between two population means

Theorem (Confidence interval of $\mu_1 - \mu_2$)

Suppose the population std. dev. σ_1, σ_2 are known.

- *If the population distributions are normal*
- *Or if the sample sizes are large enough $n_1 \geq 30, n_2 \geq 30$, then regardless of the population distributions*

We have with $(1-\alpha) \times 100\%$ confidence that

$$\mu_1 - \mu_2 \in CI_{(1-\alpha)\times 100\%}$$

$$\in \left[(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Confidence interval of difference between two population means

- What if the population std. dev. σ_1, σ_2 are unknown?
- When we have $\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$, we have

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

  Thus we get the critical value for the CI from standard normal distribution.
- If n_1, n_2 are large enough (say $n_1 \geq 30, n_2 \geq 30$), when we use s_1, s_2 instead of σ_1, σ_2 it has the same distribution, i.e.

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

  So we can still get the critical value from standard normal distribution.
- Notice that when $n_1 < 30$ or $n_2 < 30$, the distribution

Confidence interval of difference between two population means

Theorem (Confidence interval of $\mu_1 - \mu_2$)

Suppose the population std. dev.s σ_1, σ_2 are unknown.

- *If the sample sizes are large enough $n_1 \geq 30, n_2 \geq 30$, then regardless of the population distributions*

We have with $(1-\alpha) \times 100\%$ confidence that

$$\mu_1 - \mu_2 \in CI_{(1-\alpha)\times 100\%}$$

$$\in \left[(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

Sample distribution of sample proportion

- Suppose we are interested in a characteristic, and want to know the **population proportion**, say p, of individuals having this characteristic for the population. E.g. what if we want to know the proportion of humans with blue eyes?

Sample distribution of sample proportion

- Suppose we are interested in a characteristic, and want to know the **population proportion**, say p, of individuals having this characteristic for the population. E.g. what if we want to know the proportion of humans with blue eyes?
- We would take a sample, and measure the **sample proportion**, say $\hat{p}$, of individuals having the characteristic we are interested in for the sample.
- $\hat{p}$ is the estimator of p.
- What is the distribution of $\hat{p}$?

Sample distribution of sample proportion

- Let X be the number of individuals having the characteristic in the sample. Then the sample proportion

$$\hat{p} = \frac{X}{n}$$

When $n \ll N$, $X \sim \text{Binomial}(n, p)$.

- We have

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

and

$$V(\hat{p}) = V\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 V(X) = \left(\frac{1}{n}\right)^2 npq = \frac{pq}{n}$$

Sample distribution of sample proportion

- Recall we saw previously that one application of CLT is estimating the Binomial distribution: when $np > 5, nq > 5$ we have

$$X \sim \text{Binomial}(n, p) \overset{\text{approx.}}{\sim} \mathcal{N}(np, \sqrt{npq})$$

- Then we have

$$\hat{p} \overset{\text{approx.}}{\sim} \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$$

Sampling distribution of sample proportion

Theorem (Sampling distribution of $\hat{p}$)

We gather a random sample of size n from the population of size N ($n \ll N$). The sampling distribution of the sample proportion $\hat{p}$ has the following properties:

1. *Mean:* $E(\hat{p}) = p$
2. *Standard deviation (standard error):* $SE(\hat{p}) = SD(\hat{p}) = \sqrt{\frac{pq}{n}}$
3. *Shape: If* $np > 5, nq > 5$, *we have*

$$\hat{p} \overset{approx.}{\sim} \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$$

Sampling distribution of sample proportion

Example (Blue Eyes)

Suppose that 9% of humans on Earth have blue eyes. What is the probability that at least 11% in a sample of 613 individuals will have blue eye?

Sampling distribution of sample proportion

We're asked to find $P(\hat{p} \geq 0.11)$.

$$np = 613 \times 0.09 = 55.17 > 5, nq = 613 \times 0.91 = 557.83 > 5$$

Then we know $\hat{p} \overset{\text{approx.}}{\sim} \mathcal{N}(p, \sqrt{pq/n})$.

Sampling distribution of sample proportion

We're asked to find $P(\hat{p} \geq 0.11)$.

$$np = 613 \times 0.09 = 55.17 > 5, nq = 613 \times 0.91 = 557.83 > 5$$

Then we know $\hat{p} \overset{\text{approx.}}{\sim} \mathcal{N}(p, \sqrt{pq/n})$.
The mean is

$$\mu_{\hat{p}} = p = 0.09$$

and the standard error is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.09 \cdot 0.91}{613}} \approx 0.0121.$$

Sampling distribution of sample proportion

We're asked to find $P(\hat{p} \geq 0.11)$.

$$np = 613 \times 0.09 = 55.17 > 5, nq = 613 \times 0.91 = 557.83 > 5$$

Then we know $\hat{p} \overset{\text{approx.}}{\sim} \mathcal{N}(p, \sqrt{pq/n})$.
The mean is

$$\mu_{\hat{p}} = p = 0.09$$

and the standard error is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.09 \cdot 0.91}{613}} \approx 0.0121.$$

Therefore

$$P(\hat{p} \geq 0.11) \approx 1 - \Phi\left(\frac{0.11 - 0.09}{0.0121}\right) \approx 0.0564$$

Confidence interval of population proportion

- Consider the confidence interval for p
- When we want to get the confidence interval for p, we don't know the information about it. We should suppose p is unknown.
- We use $\hat{p}$ instead of p in $SE(\hat{p})$ and the condition for np, nq, because $\hat{p}$ is a good estimator of p.

Confidence interval of population proportion

Theorem (Confidence interval of p)

Suppose the population proportion p is unknown.
If $n\hat{p} > 5, n\hat{q} > 5$, we have with $(1-\alpha)\times 100\%$ confidence that

$$p \in CI_{(1-\alpha)\times 100\%} = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

Difference of sample proportions

We may also be interested in the difference in proportions of two different populations. E.g.

- In vaccine trials, we would want to know the proportion of individuals who contract the disease in the control group versus the treatment group.
- The proportion of democrat and republican voters who are in favor of a certain policy.
- The proportion of defective items manufactured in two different production lines.

Difference of sample proportions

Suppose that random samples of n_1 & n_2 have been collected from two populations with parameters p_1 & p_2, respectively. Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions of individuals meeting the criteria of interest. Then...

Sampling distribution of difference of two sample proportions

Theorem (Sampling distribution of $\hat{p}_1 - \hat{p}_2$)

Given two populations of size N_1 and N_2, we gather two independent random samples from them of size n_1 and n_2, respectively ($n_1 \ll N_1, n_2 \ll N_2$). The sampling distribution of the difference $\hat{p}_1 - \hat{p}_2$ has the following properties:

1. *Mean:* $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
2. *Standard deviation (standard error)*
   $SE(\hat{p}_1 - \hat{p}_2) = SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$
3. *Shape: If $n_1 p_1 > 5, n_1 q_1 > 5$ and $n_2 p_2 > 5, n_2 q_2 > 5$, we have*

$$\hat{p}_1 - \hat{p}_2 \overset{approx.}{\sim} \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Confidence interval of difference of two population proportions

Theorem (Confidence interval of $p_1 - p_2$)

Suppose the population proportions p_1, p_2 are unknown. If $n_1\hat{p}_1 > 5, n_1\hat{q}_1 > 5$ and $n_2\hat{p}_2 > 5, n_2\hat{q}_2 > 5$, we have with $(1-\alpha)\times 100\%$ confidence that

$$p_1 - p_2 \in CI_{(1-\alpha)\times 100\%} \in$$

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right]$$

Confidence interval of difference of two population proportions

Example (Bond Proposal Opinions)

A bond proposal will be submitted to the voters at the next local election. The majority of this bond will go to building new schools in a developing section of the city, while the remaining will go to improvements on schools in the rest of the city. A random sample sample of $n_1 = 50$ residents in the developing section and $n_2 = 100$ residents from the rest of the city. Each resident was asked weather they plan to vote for the proposal:

	developing section	rest of city
sample size	50	100
proportion in favor	0.76	0.65

Estimate the difference in the true proportions favoring the bond proposal with a 99% confidence interval.

Confidence interval of difference of two population proportions

We first check that
$n_1\hat{p}_1 = 38 > 5, n_1\hat{q}_1 = 12 > 5, n_2\hat{p}_2 = 65 > 5, n_2\hat{q}_2 = 35 > 5.$
The difference in our sample proportions are

$$\hat{p}_1 - \hat{p}_2 = 0.76 - 0.65 = 0.11$$

with an estimated standard error of

$$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(0.76)(0.24)}{50} + \frac{(0.65)(0.35)}{100}} \approx 0.0770.$$

Confidence interval of difference of two population proportions

We first check that
$n_1\hat{p}_1 = 38 > 5, n_1\hat{q}_1 = 12 > 5, n_2\hat{p}_2 = 65 > 5, n_2\hat{q}_2 = 35 > 5.$
The difference in our sample proportions are

$$\hat{p}_1 - \hat{p}_2 = 0.76 - 0.65 = 0.11$$

with an estimated standard error of

$$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(0.76)(0.24)}{50} + \frac{(0.65)(0.35)}{100}} \approx 0.0770.$$

For a 99% CI we use the critical value $z_{\alpha/2} = z_{0.005} \approx 2.58$, which gives

$$0.11 \pm (2.58)(0.0770)$$

Choosing a Sample Size

- If you want to conduct a statistical experiment in real life, you will usually have to submit an experimental design to procure funding.

Choosing a Sample Size

- If you want to conduct a statistical experiment in real life, you will usually have to submit an experimental design to procure funding.
- This leads to the question of how large of a sample size do we need?

Choosing a Sample Size

- If you want to conduct a statistical experiment in real life, you will usually have to submit an experimental design to procure funding.
- This leads to the question of how large of a sample size do we need?
- Of course, we'd like it to be as large as possible to get the best estimates. But this may not be feasible (due to time or expenses), so how can we decide the sample size before beginning the experiment?

Choosing a Sample Size

- Generally, we would pick a desired margin of error for our CI, say ε.
  So we need $ME \leq \varepsilon$

Choosing a Sample Size

- Generally, we would pick a desired margin of error for our CI, say ε.
  So we need $ME \leq \varepsilon$
- For the population mean μ, we have

$$z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \leq \varepsilon \implies n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \times \sigma^2$$

- For the population proportion p, we have

$$z_{\alpha/2} \times \sqrt{\frac{pq}{n}} \leq \varepsilon \implies n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \times pq$$

Choosing a Sample Size

- Now we know α and ε. (we know the confidence level and we want the ME of the CI$\leq \varepsilon$)
- For the population mean μ
  1. If we know σ, we can directly give the range of n as $n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \times \sigma^2$
  2. If we don't know σ, we need to estimate σ when we just know a little about the population, and then we give the range of n as $n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \times \sigma^2$
- For the population proportion p
  We don't know p (otherwise we don't need to do the experiment to estimate it)
  We need to estimate the parameter p before we know any information about the population (so it is the estimate for p for the worst case), and then we give the range of n as $n \geq \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \times pq$

Choosing a Sample Size

Parameter	Estimate	Example
σ	$\sigma = (\text{Pop. Range})/4$	Ages of college students might range from 17 to 60. Thus $\sigma \approx 10.75$
p	$p = 1/2$	In an agree or disagree poll, we would assume (half agree and half disagree with no prior knowledge of the pop.)

Rules of Thumb; Cont.

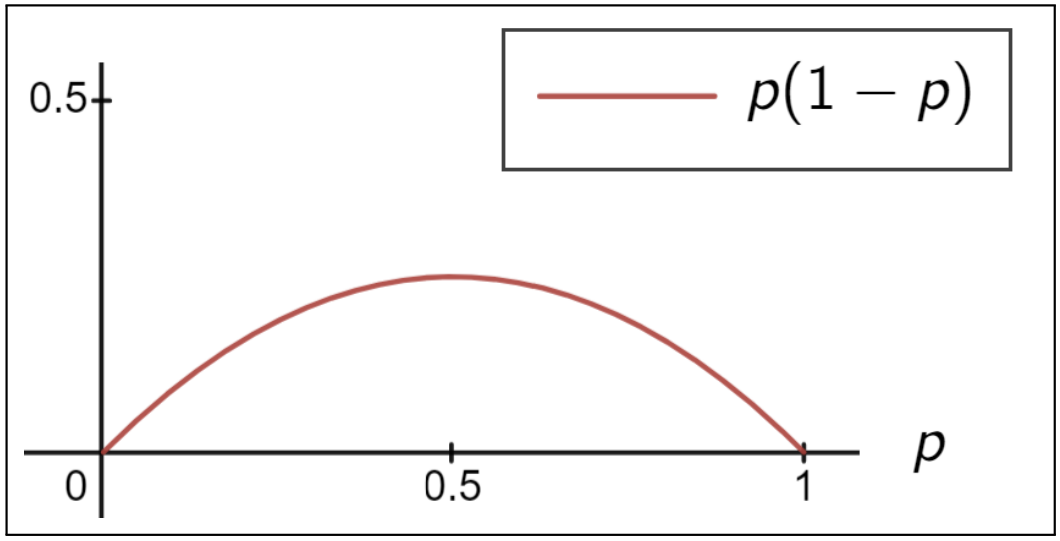
The reason for these rules is pretty simple:

1. 95% of all observations from a standard normal distribution will fall within ± 2. So if we are sampling from a $\mathcal{N}(\mu, \sigma)$ distribution (which often the case!) then $1/4$ of the range would be a pretty good guess for σ.

Rules of Thumb; Cont.

The reason for these rules is pretty simple:

1. 95% of all observations from a standard normal distribution will fall within ± 2. So if we are sampling from a $\mathcal{N}(\mu, \sigma)$ distribution (which often the case!) then $1/4$ of the range would be a pretty good guess for σ.
2. If we are trying to minimize the the margin error, then the worst case scenario (i.e. when the SE is largest) is when $p = 1/2$:



Choosing a Sample Size

Question

Suppose we want to estimate p with a 95% margin of error of <u>at most</u> 0.01. How large should we choose the sample size n?

Choosing a Sample Size

Question

Suppose we want to estimate p with a 95% margin of error of at most 0.01. How large should we choose the sample size n?

Set up equation

$$ME = 1.96\sqrt{\frac{p(1-p)}{n}} \leq 0.01$$

and we'd like to solve for n.

The difficulty is that we don't know p. We could use $\hat{p}$, but we first need n to compute $\hat{p}$ (so $\hat{p}$ depends on n), so we cannot proceed this way.

Confidence interval of population proportion

Solution:

We have

$$n \geq \left(\frac{1.96}{0.01}\right)^2 pq$$

If we plot the graph of $y = p(1-p)$ it looks like the curve, where $y = 0$ for $p = 0$ and $p = 1$, and its maximum value $y = 0.5(1-0.5)$ appears for $p = 0.5$.

This is to say: for any possible value of $p \in [0, 1]$, we have $p(1-p) \leq 0.5(1-0.5)$. Therefore if we choose n such that

$$n \geq \left(\frac{1.96}{0.01}\right)^2 0.5(1-0.5) = 9604,$$

it is enough