# Math C067 — Confidence Intervals

Richard Beigel

April 10, 2006

**Parameters vs. Statistics**

**A parameter** is a numerical property (characteristic) of a **population**. Examples:

- $\mu$ (population mean)
- $\sigma^2$ (population variance)
- $\sigma$ (population standard deviation)

A **parameter is just a number**, which may be **known or unknown**.

**A statistic** is a numerical property (characteristic) of a **sample**.

- $\bar{x}$ (sample mean — fixed sample)
- $\overline{X}$ (sample mean — random sample)
- $s^2$ (sample variance)
- $s$ (sample standard deviation)

If the sample is random, then the **statistic is a random variable**.

A statistic may depend on **known** parameters of the populatation.

**A point estimate** is a statistic that is used in order to estimate a parameter.

**Unbiased Estimator** A statistic is called an *unbiased estimator* of a population parameter if the statistic's expected value is equal to the parameter. Examples:

- The sample mean $\overline{X}$ is an unbiased estimator of the population mean $\mu$.
- The sample variance $s^2$ is an unbiased estimator of the population variance $\sigma^2$
- The sample proportion $\hat{P}$ is an unbiased estimator of the success probability $p$
- $\frac{\hat{P}(1-\hat{P})}{n-1}$ is an unbiased estimator for the variance of the sample proportion $\hat{P}$

**Biased Estimator** A statistic is called a *biased estimator* of a population parameter if the statistic's expected value is **not equal** to the parameter. (Typically they are approximately equal if the sample size $n$ is large.)

- $\frac{1}{n}\left((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right) = \frac{n-1}{n}s^2$ is a biased estimator of $\sigma^2$.
- $\frac{\hat{P}(1-\hat{P})}{n}$ is a biased estimator for the variance of the sample proportion $\hat{P}$

**Confidence Interval**

- Population random variable $X$ with mean $\mu$
- Sample mean $\overline{X}$
- We want to use $\overline{X}$ to estimate $\mu$ with 95% confidence.
- $E$ denotes **margin of error**
- $\bar{x}$ denotes the result of a particular random sample

The event that $\overline{X}$ estimates $\mu$ within a margin of error $E$ can be expressed as follows:

- $|\overline{X} - \mu| \leq E$
- $\boldsymbol{\mu - E \leq \overline{X} \leq \mu + E}$
- $\overline{X} - E \leq \mu \leq \overline{X} + E$

Those conditions are equivalent, so they have the same probability:

$$P(|\overline{X} - \mu| \leq E) = P(\mu - E \leq \overline{X} \leq \mu + E) = P(\overline{X} - E \leq \mu \leq \overline{X} + E)$$

**Definition:** If there is a 95% probability that $\overline{X}$ estimates $\mu$ with margin of error $E$ or less, i.e., if

$$P(\mu - E \leq \overline{X} \leq \mu + E) = 0.95$$

then

- $[\overline{X} - E, \overline{X} + E]$ is called a **random 95%-confidence interval** for $\mu$.
- $[\bar{x} - E, \bar{x} + E]$ is called a **95%-confidence interval** for $\mu$.
- Nothing special about 95. 99% confidence intervals, etc., defined similarly.

**Example: Finding Margin of Error** Suppose that the population random variable $X$ is normally distributed with unknown mean $\mu$ and standard deviation 4. A random sample of size 25 is used in order to estimate $\mu$. What is the margin of error for a 97% confidence interval for $\mu$?

**Solution:** Because $X$ is normally distributed, $\overline{X}$ is also normally distributed and has the same mean.

- Recall that $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$,
- so $\sigma_{\overline{X}} = 4/5 = 0.8$

We want

$$\begin{aligned} 0.97 &= P(\mu - E \leq \overline{X} \leq \mu + E) \\ 0.97 &= P(\mu - E \leq \overline{X} \leq \mu) + P(\mu \leq \overline{X} \leq \mu + E) \\ 0.97 &= 2P(\mu \leq \overline{X} \leq \mu + E) \qquad \text{by symmetry} \\ 0.97/2 &= P(\mu \leq \overline{X} \leq \mu + E) \\ 0.485 &= P(\mu \leq \overline{X} \leq \mu + E) \end{aligned}$$

- Search table A-1 for the number 0.485
- We find 0.485 in row 2.1 column 0.07
- So $E$ consists of 2.17 standard deviations
- $E = 2.17\sigma_{\overline{X}} = 2.17 \cdot 0.8 = 1.736$.

For future reference: There is no column labeled 0.485 in the Student's $t$ Table (A-2), so our only option is to search Table A-1.

**Example: Finding Confidence Level** Suppose that the population random variable $X$ is normally distributed with unknown mean $\mu$ and standard deviation 4. A random sample of size 25 is used in order to estimate $\mu$. We wish to estimate $\mu$ within a margin of error of 2. What is the corresponding confidence level?

**Solution:** Because $X$ is normally distributed, $\overline{X}$ is also normally distributed and has the same mean.

- Recall that $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$,
- so $\sigma_{\overline{X}} = 4/5 = 0.8$

As in Chapters 6 and 7, we want to find

$$P(\mu - 2 \leq \overline{X} \leq \mu + 2) = P(\mu - 2 \leq \overline{X} \leq \mu) + P(\mu \leq \overline{X} \leq \mu + 2)$$

Each subinterval consists of $\frac{2}{0.8} = 2.5$ standard deviations, so (looking in row 2.5, column 0 of Normal Distribution Table A-1)

$$P(\mu - 2 \leq \overline{X} \leq \mu + 2) \approx 0.4938 + 0.4938 = 0.9876,$$

i.e., $[\overline{X} - 2, \overline{X} + 2]$ is a 98.76% confidence interval for $\mu$.

**Example: Choosing Sample Size** Suppose that the population random variable $X$ is normally distributed with unknown mean $\mu$ and standard deviation 4. We wish to obtaina 95% confidence interval for $\mu$ with margin of error 0.5 or less. How large must the sample size be?

**Solution:** Because $X$ is normally distributed, $\overline{X}$ is also normally distributed and has the same mean.

- Recall that $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$,
- so $\sigma_{\overline{X}} = 4/\sqrt{n}$

We want

$$\begin{aligned} 0.95 &= P(\mu - 0.5 \leq \overline{X} \leq \mu + 0.5) \\ 0.95 &= P(\mu - 0.5 \leq \overline{X} \leq \mu) + P(\mu \leq \overline{X} \leq \mu + 0.5) \\ 0.95 &= 2P(\mu \leq \overline{X} \leq \mu + 0.5) \qquad \text{by symmetry} \\ 0.95/2 &= P(\mu \leq \overline{X} \leq \mu + 0.5) \qquad \text{by symmetry} \\ 0.475 &= P(\mu \leq \overline{X} \leq \mu + 0.5) \end{aligned}$$

- Search table A-1 for the number 0.475
- We find 0.475 in row 1.9 column 0.06
- So 0.5 consists of 1.96 standard deviations, i.e.,

$$\begin{aligned} 0.5 &= 1.96\sigma_{\overline{X}} \\ 0.5 &= 1.96\frac{\sigma_X}{\sqrt{n}} \\ 0.5 &= 1.96\frac{4}{\sqrt{n}} \\ \sqrt{n} &= 1.96\frac{4}{0.5} \\ \sqrt{n} &= 15.68 \\ n &= 15.68^2 \\ n &= 245.8624 \end{aligned}$$

The sample size must be an integer, so the sample size must be at least 246.

For future reference: Instead of searching the Normal Distribution Table (A-1), we could have looked in the last row (labeled $\infty$) of the Student's $t$ Distribution Table (A-2). The last entry in the column labeled 0.475 is 1.96.

**Confidence Intervals When $\sigma_X$ Is Unknown: Student's $t$ Distribution** When $\sigma_X$ is unknown, we have to estimate $\sigma_X$ using the formula for sample variance:

$$s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right)$$

and then taking the square root.

In this case it is no longer correct to use Table A-1 (Normal Distribution), and we must use Table A-2 (Student's $t$ Distribution) instead. Why? We had to divide by $\sigma$ in order to use table A-1. That was fine when $\sigma$ was just a number. But now we are dividing by the sample variance, which is itself a distribution. (It's ok if you want to think of the sample variance as a number; just remember to use Table A-2).

Table A-2 is not used like table A-1. To use Table A-2, look in the row labeled $n-1$ and the column labeled $\frac{1}{2}\gamma$ where $n$ is the sample size and $\gamma$ is the desired confidence level. If the number in that row and column is $t^*$, then your margin of error is equal to $t^*$ standard deviations.

Note: If you are using the NIST Table online, look in the column labeled $\frac{1}{2} - \frac{1}{2}\gamma$. If you are using a table with a 2-tail header, look in the column labeled $1 - \gamma$.

Requirement: Either the population random variable $X$ is normally distributed or the sample size is at least 30 ($\overline{X}$ is approximately normally distributed).

**Example: Finding Margin of Error when $\sigma$ is unknown** Suppose that the population random variable $X$ is normally distributed with unknown mean $\mu$ and unknown standard deviation $\sigma$. A random sample of size 25 is used in order to estimate $\mu$. The sample variance for the random sample is calculated to be 16.00. What is the margin of error for a 98% confidence interval for $\mu$?

**Solution:** Because $X$ is normally distributed, $\overline{X}$ is also normally distributed and has the same mean.

- $\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{16.00} = 4.00$.
- Recall that $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$,
- The sample variance is an unbiased estimator for $\sigma_X$, so we **estimate** that $\sigma_{\overline{X}}$ is $4.00/\sqrt{n} = 4.00/5 = 0.8$.
- We have $25 - 1 = 24$ degrees of freedom so we look in row 24.
- The desired confidence level $\gamma$ is 0.98. Where we look depends on the kind of $t$ table we are using
  - If we are using Table A-2 of the Students $t$ Distribution, we look in column $\frac{1}{2}\gamma = 0.49$.
  - If we are using NIST's 1-tail table of the Student's $t$ Distribution, we look in column $\frac{1}{2} - \frac{1}{2}\gamma = 0.01$.
  - If we are using a 2-tail table of the Student's $t$ Distribution we look in column $1 - \gamma = 0.02$

We find that for a confidence level of 98% with 25 samples, the margin of error consists of approximately 2.492 standard deviations, so the margin of error is approximately $2.492 \cdot 0.8 = 1.9936$.

Note: When we knew that $\sigma_X$ was 4, our margin of error was only 1.736. The increase in the margin of error (from 1.736 to 1.9936) is due to our uncertainty in estimating the standard deviation.

**Example 2: Election Prediction Based on a Small Sample** You have been hired to predict your home town election result. Because of budget constraints you are only able to poll 50 voters. Suppose that 30 voters in your sample prefer Candidate C.

1. Predict the percentage of voters in town who prefer Candidate C.
2. What is the margin of error for a 99% confidence level?

**Solution:**

- Let $\hat{P}$ denote the sampling proportion for $n = 50$ samples, where each sample has unknown success probability $p$.
- The proportion favoring Candidate C in our sample is $\hat{p} = 0.6$.
- Because $\hat{P}$ is an unbiased estimator for $p$, $\hat{p}$ is our estimate for $p$
- i.e., we predict that 60% of the population will favor Candidate C.
- Because the **sample** standard deviation $\frac{\hat{P}(1-\hat{P})}{n-1}$ is an unbiased estimator for $\sigma^2_{\hat{P}}$, $\sqrt{\frac{\hat{P}(1-\hat{P})}{n-1}}$ is our estimate for $\sigma_{\hat{P}}$,
- i.e., we estimate $\sigma_{\hat{P}}$ as $\sqrt{\frac{0.6 \cdot 0.4}{49}} \approx 0.07$.
- We need to find the number $E$ such that $P(p - E \leq \hat{P} \leq p + E) = 0.99$
- We have $50 - 1 = 49$ degrees of freedom so we look in row 49 of the Student's $t$ Table.
- The desired confidence level $\gamma$ is 0.99. Where we look depends on the kind of $t$ table we are using
  - If we are using Table A-2 of the Students $t$ Distribution, we look in column $\frac{1}{2}\gamma = 0.495$.
  - If we are using NIST's 1-tail table of the Student's $t$ Distribution, we look in column $\frac{1}{2} - \frac{1}{2}\gamma = 0.005$.
  - If we are using a 2-tail table of the Student's $t$ Distribution we look in column $1 - \gamma = 0.01$

  We find that for a confidence level of 99% with 49 degrees of freedom, $E$ consists of approximately 2.68 standard deviations.
- So $E \approx 2.68 \cdot 0.07 = 0.1876$, i.e., the margin of error is approximately 18.76%.

**Example 2: Election Prediction Based on a Large Sample** Candidate C received only 49% of the votes and lost your home town election, and your local newspaper editor, who didn't seem to notice that this result was within your margin of error, fired you. Fortunately for you, NBC was looking for a statistician who understands margin of error. NBC hired you and gave you a huge polling budget, so you are able to poll 5000 voters for a statewide primary election. Suppose that 3000 voters in your sample prefer Candidate Q.

1. Predict the percentage of the population whoo prefer Candidate Q.
2. What is the margin of error for a 99% confidence level?

**Solution:**

- Let $\hat{P}$ denote the sampling proportion for $n = 5000$ samples, where each sample has unknown success probability $p$.
- The proportion favoring Candidate Q in our sample is $\hat{p} = 0.6$.
- Because $\hat{P}$ is an unbiased estimator for $p$, $\hat{p}$ is our estimate for $p$
- i.e., we predict that 60% of the population will favor Candidate Q.
- Because $\frac{\hat{P}(1-\hat{P})}{n-1}$ is an unbiased estimator $\sigma^2_{\hat{P}}$, $\sqrt{\frac{\hat{P}(1-\hat{P})}{n-1}}$ is our estimate for $\sigma_{\hat{P}}$,
- i.e., we estimate $\sigma_{\hat{P}}$ as $\sqrt{\frac{0.6 \cdot 0.4}{4999}} \approx 0.006929$.
- We need to find a number $E$ such that $P(p - E \leq \hat{P} \leq p + E) = 0.99$
- We have $5000 - 1 = 4999$ degrees of freedom. The Student's $t$ Table doesn't go that far, because for really large $n$ the Student's $t$ Distribution looks like a normal distribution, we look in the row labeled $\infty$.
- The desired confidence level $\gamma$ is 0.99. Where we look depends on the kind of $t$ table we are using
  - If we are using Table A-2 of the Students $t$ Distribution, we look in column $\frac{1}{2}\gamma = 0.495$.
  - If we are using NIST's 1-tail table of the Student's $t$ Distribution, we look in column $\frac{1}{2} - \frac{1}{2}\gamma = 0.005$.
  - If we are using a 2-tail table of the Student's $t$ Distribution we look in column $1 - \gamma = 0.01$
  - If there is no row labeled $\infty$ then **search** for $\frac{1}{2}\gamma = 0.495$ in a Normal Distribution Table (A-1).

  We find that for a confidence level of 99% with mucho degrees of freedom, $E$ consists of approximately 2.58 standard deviations.
- So $E \approx 2.58 \cdot 0.006929 \approx 0.01788$, i.e., the margin of error is approximately 1.788%

**Example 3: Choosing Sample Size for Election Prediction** Candidate Q won the primary 59% to 41%. NBC was very happy with your work. They increased your polling budget and requested that you predict the results of the general election with a 1% margin of error and 99% confidence. How big must your sample be?

**Solution:**

- Let $\hat{P}$ denote the sampling proportion for $n$ samples, where each sample has unknown success probability $p$.
- The standard deviation of $\hat{P}$ is $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$.
- We don't know what $p$ is, but we are guaranteed that

$$\begin{aligned} \sigma_{\hat{P}} &\leq \sqrt{\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{n}} \\ &= \sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{n}} \\ &= \frac{1}{2}\frac{1}{\sqrt{n}} \end{aligned}$$

- We need to find a number $E$ such that $P(p - E \leq \hat{P} \leq p + E) = 0.99$
- As before, $E$ consists of approximately 2.58 standard deviations, so

$$\begin{aligned} E &= 2.58\sigma_{\hat{P}} \\ &\leq 2.58 \cdot \frac{1}{2}\frac{1}{\sqrt{n}} \\ &= \frac{1.29}{\sqrt{n}} \end{aligned}$$

- i.e., $E \leq \dfrac{1.29}{\sqrt{n}}$.
- For a 1% margin of error, we need $E \leq 0.01$, so we need to make

$$\begin{aligned} \frac{1.29}{\sqrt{n}} &\leq 0.01 \\ \frac{1.29}{0.01} &\leq \sqrt{n} \\ 129 &\leq \sqrt{n} \\ 129^2 &\leq n \\ 16641 &\leq n \end{aligned}$$

- In other words a sample of size 16641 (or larger) will provide a 1%-or-smaller margin of error with 99% confidence. (The actual margin of error will be smaller if $p \neq \frac{1}{2}$.)