AP STATISTICS
TOPIC VIII: ESTIMATION (DRAFT)

PAUL L. BAILEY

1. Confidence Intervals

1.1. **Definition of Confidence Interval.** Let $\gamma \in \mathbb{R}$ and let $c \in [0, 1]$. Here, γ is a value we wish to estimate, and c is a probability.

A *confidence interval of level c for γ* (aka *c-confidence interval*) is a bounded open interval I such that $P(\gamma \in I) = c$.

A bounded open interval is of the form $I = (a, b)$. Let $g = \dfrac{a+b}{2}$ be the midpoint of this interval, and let $E = b - g$. Then $I = (g - E, g + E)$. We call I a *symmetric open interval about g of radius E*. Think of g as our estimate for γ, and E as the tolerance for error in the estimate. Then c is the probability that the actual value γ is within the error tolerance of the estimate.

1.2. **Parameters and Statistics Revisited.** A *parameter* is a number computed using the entire population. A *statistic* is a number computed using a sample of the population. The statistics are computed using the same algorithms as the parameters, just on smaller sets. We have seen the following examples of parameters and corresponding statistics.

Name	Parameter	Statistic
Mean	μ	$\overline{x}$
Variation	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\widehat{p}$
Generic	γ	g

A *point estimate* of a population parameter is an estimate of the parameter using a corresponding statistic. The *margin of error* of the statistic g used as an estimate for the parameter γ is

$$|g - \gamma|.$$

An *error tolerance*, denoted E, is a measure of how small we wish $|g - \gamma|$ to be; that is, we want $|g - \gamma| < E$. Note that

$$|g - \gamma| < E \quad \Leftrightarrow \quad g - E < \gamma < g + E \quad \Leftrightarrow \quad \gamma \in (g - E, g + E).$$

How do we find a confidence interval? We seek the error tolerance E such that

$$P(g - E < \gamma < g + E) = c.$$

For estimating the mean μ of a population from a sample mean $\overline{x}$, we have the tools to do this in the case that the population is approximately normal and the standard deviation σ is known.

Date: November 20, 2023.

2. Point Estimate for the Mean

2.1. **Development of the Error Tolerance.** Consider a population with mean μ and standard deviation σ. We take a sample of size n. The mean of the sample is $\overline{x}$ and the standard deviation is s. We view $\overline{x}$ as an estimate for μ.

The *margin of error* of this point estimate for the mean is

$$|\overline{x} - \mu|.$$

We wish this estimate to be no worse than our error tolerance E, so that $|\overline{x}-\mu| < E$. We have

$$|\overline{x} - \mu| < E \quad \Leftrightarrow \quad \overline{x} - E < \mu < \overline{x} + E \quad \Leftrightarrow \quad \mu \in (\overline{x} - E, \overline{x} + E).$$

Let $c \in [0, 1]$. A *confidence interval for μ at level c based on $\overline{x}$* is a symmetric open interval about $\overline{x}$ of radius E such that

$$P(\overline{x} - E < \mu < \overline{x} + E) = c.$$

If the population has a normal distribution or if n is large, then $\overline{x}$ has an approximately normal distribution. In order to compute the confidence interval, we need the inverse cumulative density function. Since calculators with this functionality are a relatively recent technological development, it is traditional to begin with the standard normal distribution, or z-score.

We wish to find a interval symmetric about zero such that the probability that a random z-value is in this interval is c; that is, we want to a number z_c so that the area under the curve of the standard normal distribution from $-z_c$ to z_c is c. In notation, we want z_c so that

$$P(-z_c < z < z_c) = \int_{-z_c}^{z_c} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = c.$$

For such a z_c, we have $P(z > z_c) = \dfrac{1-c}{2}$, so $P(z < z_c) = 1 - \dfrac{1-c}{2} = \dfrac{1+c}{2}$.

We define the *critical value* of z for c to be the positive real number z_c such that

$$P(z < z_c) = \frac{1+c}{2}.$$

The z-score that corresponds to our point estimate $\overline{x}$ is given by

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}.$$

Thus

$$\begin{aligned} -z_c < z < z_c \quad &\Leftrightarrow \quad -z_c < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_c \\ &\Leftrightarrow \quad -z_c \frac{\sigma}{\sqrt{n}} < \overline{x} - \mu < z_c \frac{\sigma}{\sqrt{n}} \\ &\Leftrightarrow \quad \overline{x} - z_c \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_c \frac{\sigma}{\sqrt{n}} \end{aligned}$$

From this, we see that if we set the error tolerance to

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

then we have

$$P(\overline{x} - E < \mu < \overline{x} + E) = c.$$

Thus $(\overline{x} - E, \overline{x} + E)$ is the c-confidence interval for $\overline{x}$ as a point estimate for μ.

2.2. **Nuts and Bolts.** We discuss these computations within the context of a problem. We assume that the reader has a scientific calculator and a z-table.

Problem 1 (Brase §8.1 # 11)**. (Zoology: Hummingbirds)**
Allen's hummingbird (Selasphorus sasin) has been studied by zoologist Bill Alther. A small group of 15 Allen's hummingbirds has been under study in Arizona. The average weight for these birds is $\overline{x} = 3.15$g. Based on previous studies, we can assume that the weights of Allen's hummingbirds have a normal distribution, with $\sigma = 0.33$g.

- **(a)** Find an 80% confidence interval for the average weights of Allen's hummingbirds in the study region. What is the margin of error?
- **(b)** What conditions are necessary for your calculations?
- **(c)** Give a brief interpretation of your results in the context of this problem.
- **(d)** Find the sample size necessary for an 80% confidence level with a maximal error of estimate $E = 0.08$ for the mean weights of the hummingbirds.

Solution. We have $\sigma = 0.33$, $n = 15$, and $\overline{x} = 3.15$. Note that μ is unknown.

- **(a)** We have $c = 0.8$, so we want to find z_c such that $P(z < z_c) = \dfrac{1+c}{2} = 0.9$. We look up 0.9 on a z-table and find the $P(z < 1.28) \approx 0.8997$ and $P(z < 1.29) < 0.9015$. We take the *more conservative* value, which is $z_c = 1.28$.

  Now $E = z_c \dfrac{\sigma}{\sqrt{n}} = (1.28)\dfrac{0.33}{\sqrt{15}} = 0.109$. We compute $\overline{x} - E = 3.15 - 0.109 = 3.041$ and $\overline{x} + E = 3.15 + 0.109 = 3.259$. The 80% confidence interval is $(3.041, 3.259)$, which is to say that

  $$P(\mu \in (3.041, 3, 259)) \geq 80\%.$$

  WARNING: the margin of error is $|\overline{x} - \mu|$. The book says the answer to "what is the margin of error" is 0.11, but this is wrong. We do not know μ, so we do not know the margin of error. We know that the maximal margin of error is $E = 0.109$ at an 80% level of confidence.
- **(b)** The computation is valid, because the distribution is approximately normal and σ is known.
- **(c)** Our conclusion is that $P(\mu \in (3.041, 3, 259)) \geq 80$.
- **(d)** To address this part, we wish to solve the equation $E = z_c \dfrac{\sigma}{\sqrt{n}}$ for n. We obtain

  $$n = \left(\frac{z_c \sigma}{E}\right)^2 = \left(\frac{(1.28)(0.33)}{0.08}\right)^2 = 27.878.$$

  Since n is an integer, we take the conservative approach and round up to $n = 28$.

  This means that if we wish an error tolerance of 0.08 at the 80% confidence level, we need a sample size of at least $n = 28$ hummingbirds.

□

Department of Mathematics, Paragon Science Academy
Email address: pbailey@sonoranschools.org