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: December 4, 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-score 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.

□

3. Student t-distribution

3.1. Cultural Background and Formal Definition. The *gamma function*, considered for nonnegative real numbers, is defined as

$$\Gamma : [0,\infty) \to \mathbb{R} \quad \text{given by } \Gamma(x) = \int_0^\infty t^{x-1}e^{-x}\,dt.$$

This function is pivotal in analytic number theory, as it satisfies the recurrence relation

- $\Gamma(1) = 1$
- $\Gamma(x+1) = x\Gamma(x)$

This can be shown by using integration by parts. From this, it follows that the gamma function is a continuous extension of factorial; that is,

$$\Gamma(n) = (n-1)! \quad \text{for all} \quad n \in \mathbb{N}.$$

The *Student t-distribution* has probability density function

$$f_n(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

It is the sampling distribution of the statistics

$$t = \frac{\overline{x}-\mu}{\sqrt{s^2/(n-1)}},$$

where

$$\overline{x} = \sum_{i=1}^n x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i-\mu)^2.$$

We may compute confidence intervals using a table of values for the t-distribution. Let $c \in [0,1]$ be a probability. Define the *critical value* of t for c to be the positive real number t_c such that

$$P(t < t_c) = \frac{1+c}{2}.$$

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

$$t = \frac{\overline{x}-\mu}{s/\sqrt{n}}.$$

As before,

$$-t_c < t < t_c \quad \Leftrightarrow \quad -t_c < \frac{\overline{x}-\mu}{s/\sqrt{n}} < t_c \quad \Leftrightarrow \quad \overline{x} - t_c\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + t_c\frac{s}{\sqrt{n}}.$$

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

$$E = t_c\frac{s}{\sqrt{n}}$$

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.

Problem 2 (Brase §8.2 # 11)**. (Archaeology: Tree Rings)**
At Burnt Mesa Pueblo, the method of tree ring dating gave the following years A.D. for an archaeological excavation site:

```
1189   1271   1267   1272   1268   1316   1275   1317   1275
```

(a) Use a calculator with mean and standard deviation keys to verify that the sample mean year is 1272, with sample standard deviation 37 years.
(b) Find a 90% confidence interval for the mean of all tree ring dates from this archaeological site.

Solution. Use a scientific calculator.

(a) Use these formulas

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ s &= \sqrt{x} \end{aligned}$$

and you will get

$$\begin{aligned} n &= 9 \\ \overline{x} &\approx 1272 \\ s &\approx 37 \end{aligned}$$

(b) The right tail area is

$$\frac{1-c}{2} = \frac{1-0.9}{2} = 0.05.$$

The degrees of freedom are

$$d.f. = n - 1 = 8.$$

In a t-table, use this information to look up that $t_c = t_{0.05} = 1.860$. Then

$$E = t_c\frac{s}{\sqrt{n}} = \frac{(1.86)(37)}{3} = 22.94 \approx 23.$$

Thus

$$\overline{x} - E = 1249 \quad \text{and } \overline{x} + E = 1295,$$

so the confidence interval is

$$I_c = (1249, 1295).$$

□

4. Estimating Proportion

Consider population S and a subset R. The elements of R are considered to be those elements of S which have a property. That is,

$$R = \{s \in S \mid s \text{ has the given property}\}.$$

The proportion of the population with this property is

$$p = \frac{|R|}{|S|}.$$

Let $q = 1 - p$.

We view p as a population parameter. Let $n \in \mathbb{Z}$ with $n \geq 2$, and let T be a randomly chosen subset of S of size n. Let $r = |\{t \in T \mid t \in R\}|$, and set

$$\widehat{p} = \frac{r}{n}.$$

Now $\widehat{p}$ is the statistic that corresponds to parameter p.

If we let T range over the collection of all subsets of S of size n, then $\widehat{p}$ becomes a probability distribution which is approximately normal if $np > 5$ and $nq > 5$. We have seen that the mean of the $\widehat{p}$ distribution is $\mu = \mu_{\widehat{p}} = p$ and $\sigma = \sigma_{\widehat{p}} = \sqrt{\dfrac{pq}{n}}$.

Suppose p is unknown, and we are interested in estimating it. Select a random sample and let $\widehat{p}$ be the proportionality statistic for that sample. We would like to compute the confidence interval for $\widehat{p}$ as an estimate for p. We convert $\widehat{p}$ to a z-score via

$$z = \frac{\widehat{p} - \mu}{\sigma}.$$

Let z_c be the critical z value at the c confidence level. Now

$$-z_c < z < z_c \Leftrightarrow -z_c < \frac{\widehat{p} - \mu}{\sigma} < z_c \Leftrightarrow \widehat{p} - z_c\sqrt{\frac{pq}{n}} < p < \widehat{p} + z_c\sqrt{\frac{pq}{n}}.$$

Let

$$E = z_c\sqrt{\frac{pq}{n}},$$

and set the confidence interval to be

$$I_c = (\widehat{p} - E, \widehat{p} + E)$$

so that

$$P(p \in I_c) = c.$$

In practice, getting equality above is impractical; we want to know that

$$P(p \in I_c) \geq c.$$

Problem 3. (Unicorn Horn Shape)
Hagrid inspected a sample of 40 unicorns from the Forbidden Forest, and 5 were determined to have bent horns.

(a) Find an 87% confidence interval for the proportion of unicorns with bent horns.
(b) Discuss whether the sample was of sufficient size.

Solution. Note that $\widehat{p} = \dfrac{8}{40} = 0.2$. Set $\widehat{q} = 1 - \widehat{p} = 0.8$. We begin with the assumption that $\widehat{p}$ effectively approximates p. Then

$$np \approx n\widehat{p} = 8 > 5 \quad \text{and } nq \approx n\widehat{q} = 24 > 5,$$

so under this assumption, the $\widehat{p}$ distribution is approximately normal.

(a) Look up $z_c = z_{0.87} = 1.12$. Set

$$E = z_c\sqrt{\frac{pq}{n}} = (1.12)\sqrt{\frac{(.2)(.8)}{40}} = 0.07.$$

Compute $\widehat{p} - E = 0.13$ and $\widehat{p} + E = 0.27$. Thus the confidence interval is

$$I_c = (0.13, 0.27).$$

Hagrid concludes with 87% confidence that between 13% and 27% of the unicorns in the Forbidden Forest have bent horns.

(b) Hagrid is skeptical of this result; he suspects that only 10% of unicorns have bent horns. He checks his sample size by solving

$$n = \frac{z_c^2 pq}{E^2} = \frac{(1.12)^2(0.1)(0.9)}{(0.87)^2} = 14.9,$$

which he rounds up to $n = 15$. So he should have used a sample of size at least 15.

Still skeptical, he realizes that the most conservative estimate for sample size comes from setting $p = q = 0.5$, so that n is at least

$$n = \frac{z_c^2 pq}{E^2} = \frac{(1.12)^2(0.5)(0.5)}{(0.87)^2} = 41.4,$$

which he rounds up to $n = 42$. Well that was codswallop! The sample wasn't big enough after all!

□

5. Estimating Differences

5.1. **Review of the General Picture.** Let γ be a parameter and let g be the corresponding statistic.

Let $c \in [0,1]$ be a confidence level for estimating γ using g. We wish to find E such that
$$P(|g-\gamma| < E) = c.$$
The *error tolerance* is the smallest real number E satisfying the equation above, and we see that
$$|g-\gamma| < E \quad \Leftrightarrow \quad g-E < \gamma < g+E \quad \Leftrightarrow \quad \gamma \in (g-E, g+E).$$
The *confidence interval at level* c is
$$I_c = (g-E, g+E).$$

If γ comes from a normal distribution, we can use a z-table to compute E.

We view g as a sampling distribution, which is approximately normal if γ comes from an approximately normal distribution, or if n is large. The mean and standard deviation of the sampling distribution are denoted μ_g and σ_g; one would expect that, $\mu_g = \gamma$.

We convert g into a z-score by
$$z = \frac{g-\gamma}{\sigma_g}.$$

The *critical value of* z *at the* c *confidence level* is the unique real number z_c such that
$$P(|z| < z_c) = c.$$
This equates to
$$P(z < z_c) = \frac{1+c}{2}.$$
We can look up this z_c in a table, or use a calculator's inverse cumulative normal distribution function.

To find the confidence interval, we note that
$$-z_c < z < z_c \Leftrightarrow -z_c \frac{g-\gamma}{\sigma_g} < z_c \Leftrightarrow g - z_c\sigma_g < \gamma < g + z_c\sigma_g.$$
Set the error tolerance to be
$$E = z_c\sigma_g$$
and the confidence interval to be
$$I_c = (g-E, g_E)$$
so that
$$P(\gamma \in I_c) \geq c.$$

5.2. **Two Populations.** Suppose that γ_1 and γ_2 are parameters with corresponding statistics g_1 and g_2. We wish to determine how close γ_1 is to γ_2.

Let $\gamma = \gamma_1 - \gamma_2$ and let $g = g_1 - g_2$. We wish to use g to estimate γ. Thus we seek the error tolerance E such that

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

Recall that if X and Y are random variables and $a, b \in \mathbb{R}$, then the expectation and variance of their linear combinations satisfies

- $E(aX + bY) = aE(X) + bE(Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y)$

The mean of the g distribution is

$$\mu_g = g_1 - g_2.$$

The standard deviation is

$$\sigma_g = \sqrt{\sigma_{g_1}^2 + \sigma_{g_2}^2}.$$

5.3. **Approximating the Difference of Means.** Consider two normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2.

We wish to estimate $\gamma = \mu_1 - \mu_2$. The corresponding sampling distribution is $g = \overline{x}_1 - \overline{x}_2$ with sample sizes n_1 and n_2, respectively. We have

$$\mu_{\overline{g}} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_g = \sqrt{\sigma_{\overline{x}_1}^2 + \sigma_{\overline{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}.$$

Thus

$$E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Problem 4. (Unicorn Horn Length)
Hagrid noticed several female unicorns with unusually long horns, and he wished to determine if unicorn horn lengths varied by gender. Let μ_1 denote the average horn length of males, and μ_2 the average horn length of females. Let $\sigma_1 = 2.30$ inches and $\sigma_2 = 1.75$ inches be the corresponding standard deviations.

A sample of $n_1 = 10$ males and $n_2 = 12$ females had respective sample average horn lengths of $\overline{x}_1 = 7.35$ and $\overline{x}_2 = 6.91$ inches, respectively.

(a) Suppose that $\sigma_1 = 2.30$ inches and $\sigma_2 = 1.75$ inches are the corresponding standard deviations. Find an 80% confidence interval for the difference $\mu_1 - \mu_2$.

(b) Suppose that σ_1 and σ_2 are unknown, and that the sample standard deviations are $s_1 = 2.12$ and $s_2 = 1.55$. Find an 80% confidence interval for the difference $\mu_1 - \mu_2$.

Solution. Let $\gamma = \mu_1 - \mu_2$ and let $g = \overline{x}_1 - \overline{x}_2 = 7.35 - 6.91 = 0.440$.

(a) Since σ_1 and σ_2 are known, we use the z-distribution. We have $n_1 = 10$, $n_2 = 12$, $\sigma_1 = 2.30$, and $\sigma_2 = 1.75$. Look up $z_c = z_{0.8} = 1.28$. Thus

$$E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (1.28)\sqrt{\frac{(2.30)^2}{10} + \frac{(1.75)^2}{12}} = 1.13$$

The $c = 80\%$ confidence interval is

$$I_c = (g - E, g + E) = (-0.69, 1.57).$$

Since $0 \in I_c$, at the 80% confidence level, Hagrid cannot rule out that horn length differs by gender.

(a) Since σ_1 and σ_2 are unknown, we use the t-distribution together with s_1 and s_2 to estimate the population standard deviations. We have $n_1 = 10$, $n_2 = 12$, $s_1 = 2.42$, and $\sigma_2 = 1.55$. The degrees of freedom are $d.f. = \min\{n_1, n_2\} - 1 = 9$. Look up $t_c = t_{0.8} = 1.38$. Thus

$$E = t_c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (1.38)\sqrt{\frac{(2.42)^2}{10} + \frac{(1.55)^2}{12}} = 1.22$$

The $c = 80\%$ confidence interval is

$$I_c = (g - E, g + E) = (-0.78, 1.66).$$

Since $0 \in I_c$, at the 80% confidence level, Hagrid cannot rule out that horn length differs by gender.

□

5.4. **Approximating the Difference of Proportions.** Let p_1 and p_2 be proportions from a pair of populations, and let $\widehat{p}_1$ and $\widehat{p}_2$ be the proportions of samples from these populations of sample size n_1 and n_2, respectively.

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