ESTIMATION

We are interested in using the values of a random sample $X_1, X_2, \ldots, X_n$ to estimate the values of population parameters (usually μ or σ^2, but also including other parameters). At this point, we have mostly talked about using the sample mean $\overline{X}$ as an estimate for the population μ (this includes using a sample proportion to estimate a population proportion). For example, if you wanted to know the mean age of the trees in a forest, you could select 100 random trees, determine their ages, then use the mean age of trees in your sample as an estimate for the mean age of trees in the whole forest.

Definition. We call the sample statistic $\hat{\Theta}$ (a random variable determined by the values of a random sample of size n) an **estimator** of the population parameter θ if the value of $\hat{\Theta}$ will be used as an estimate of θ.

(1) If $E(\hat{\Theta}) = \theta$, then we call $\hat{\Theta}$ an **unbiased estimator**.

(2) If for any $c > 0$, $\lim_{n\to\infty} P(|\hat{\Theta} - \theta| < c) = 1$, then we call $\hat{\theta}$ a **consistent estimator**.

Example. Suppose we know that population is uniformly continuously distributed on the interval $[0, \beta]$, but we don't know β. Let $\overline{X}$ be the mean of a random sample of size n. We know that $E(\overline{X}) = \frac{0+\beta}{2} = \frac{\beta}{2}$. It then follows that $E(2\overline{X}) = \beta$. Thus $\hat{B} = 2\overline{X}$ is an unbiased estimator of the population parameter β. This unbiased estimator is based on the sample mean $\overline{X}$.

1. Keep working with the population in the example. Let $X_1, X_2, \ldots, X_{10}$ be a random sample and let $\hat{X}$ be the maximum of the sample. Our goal is to find an unbiased estimator for β based on $\hat{X}$. The CDF of the population is $F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \frac{x}{\beta} & \text{if } 0 < x < \beta \\ 1 & \text{if } \beta \le x \end{cases}$

a) Find the CDF of $\hat{X}$. Hint: $\hat{X} \le x$ if and only if $X_1 \le x$ and $X_2 \le x$ and $\ldots$ and $X_{10} \le x$.
b) Differentiate the CDF to find a PDF for $\hat{X}$
c) Use the PDF to calculate $E(\hat{X})$
d) Find a number c such that $E(c\hat{X}) = \beta$

Your unbiased estimator for β is $c\hat{X}$.

CDF of $\hat{X}$: $G(x) = P(\hat{X} \le x) = P(X_1 \le x \text{ and } X_2 \le x \text{ and} \cdots X_{10} \le x)$

$= P(X_1 \le x)P(X_2 \le x) \cdots P(X_{10} \le x)$ independence

$= \left(\frac{x}{\beta}\right)^{10}$ if $0 < x < \beta$ identical dist

PDF of $\hat{X}$: $g(x) = G'(x) = \frac{10}{\beta^{10}} x^9$ if $0 < x < \beta$

$E(\hat{X}) = \int_0^\beta x \left(\frac{10}{\beta^{10}} x^9\right) dx = \int_0^\beta \frac{10}{\beta^{10}} x^{10} dx = \frac{10}{\beta^{10}} \frac{x^{11}}{11} \Big|_0^\beta = \frac{10}{11}\beta$

$\beta = E(c\hat{X}) = cE(\hat{X}) = c\frac{10}{11}\beta \Rightarrow c = \frac{11}{10}$

Estimator $\frac{11}{10}\hat{X}$

Date: March 27, 2023.

I collected a random sample from a population uniformly continuously distributed on $[0, \beta]$ and got the following values:

$$\begin{matrix} 0.8023426 & 1.1194331 & 1.5173202 & 1.5353951 & 2.1296237 \\ 2.5989737 & 3.4224997 & 3.7957449 & 4.8618741 & 5.0795233 \end{matrix}$$

This means $2\overline{x} = 5.372546$ and $1.1\hat{x} = 5.587476$ (I'm using lowercase letters here because I have actual values for the sample statistics and I'm no longer thinking of them as random variables). We now have the values of two unbiased estimators for β. Which should we actually use? Is one of the estimators more reliable than the other? There's a general principle that applies: **when given the choice between unbiased estimators, you should always choose the estimator with smaller variance.**

2. Our goal now is to determine which of the estimators has a smaller variance.

a) Use the theorems of the last worksheet to calculate $\text{Var}(2\overline{X})$ (it will help to know that the population variance is $\sigma^2 = \frac{\beta^2}{12}$).

b) You'll have to calculate the variance of the other estimator by hand. You already know the expected value of $\hat{X}$, so it makes sense to use the formula $\text{Var}(\hat{X}) = E\left(\hat{X}^2\right) - \left[E(\hat{X})\right]^2$.

Which estimator has a smaller variance?

$$\text{Var}(2\overline{X}) = 4\text{Var}(\overline{X}) = 4\left(\frac{\beta^2/12}{10}\right) = \frac{\beta^2}{30}$$

$$\text{Var}\left(\frac{11}{10}\hat{X}\right) = \left(\frac{11}{10}\right)^2 \text{Var}(\hat{X}) = \left(\frac{11}{10}\right)^2 \left[\frac{10}{12(11^2)}\beta^2\right] = \frac{\beta^2}{120}$$

$$\text{Var}(\hat{X}) = E(\hat{X}^2) - [E(X)]^2 = \frac{10}{12}\beta^2 - \left(\frac{10}{11}\beta\right)^2 = \frac{10[11^2 - 10(12)]}{12(11^2)}\beta^2 = \frac{10}{12(11^2)}\beta^2$$

$$E(\hat{X}^2) = \int_0^\beta x^2\left(\frac{10}{\beta^{10}}x^9\right)dx = \int_0^\beta \frac{10}{\beta^{10}}x^{11}\,dx = \frac{10}{12}\beta^2$$

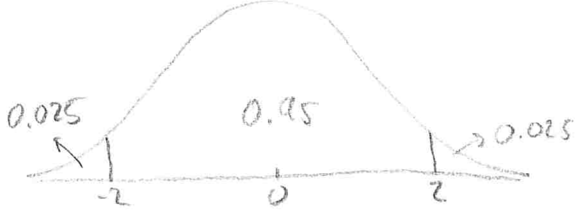
$\frac{11}{10}\hat{X}$ has a smaller variance

What we have been doing so far is called **point estimation**: we are using samples to make guesses about the values of population parameters. However, we know that our guesses are correct with probability 0 (in the case of continuous distributions). It makes sense to ask if we could come up with a range of values that is likely to contain the population parameter.[1] This is called **interval estimation**.

3. Let Z be a standard normal random variable. Find a number z such that $P(-z < Z < z) = 0.95$. The R command `qnorm(p)` returns the $100p^{\text{th}}$ percentile of the standard normal distribution; you may want to use this commmand to solve this problem.

$z = \text{qnorm}(0.975) \approx 1.959964$, often rounded to 1.96

$P(-1.96 < Z < 1.96) = 0.95$



4. Let $\overline{X}$ be the mean of a random sample of size 100 from a population with mean μ and standard deviation $\sigma = 2$ (note that this is the population standard deviation, not the standard deviation of $\overline{X}$). Substitute $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ into the expression in problem 3, then isolate μ to fill in the blanks:

$$P(\overline{X} - \underline{\qquad} < \mu < \overline{X} + \underline{\qquad}) = 0.95$$

$$-1.96 < \frac{\bar{X}-\mu}{2/10} < 1.96$$

$$-1.96(0.2) < \bar{X}-\mu < 1.96(0.2)$$

$$-\bar{X} - 0.392 < -\mu < -\bar{X} + 0.392$$

$$\bar{X} + 0.392 > \mu > \bar{X} - 0.392$$

$$P(\bar{X} - 0.392 < \mu < \bar{X} + 0.392) = 0.95$$

5. Samples are taken and you find $\overline{x} = 7.7672$. Substitute this value in for $\overline{X}$ in your solution to the previous problem to find the **95% confidence interval** for the population mean μ.

$7.7672 - 0.392 = 7.3752$

$7.7672 + 0.392 = 8.1592$

[1] We'll have to be careful with the phrase "likely to contain the population parameter." Problems 6 and 7 deal with this.

6. What's wrong with the expression $P(7.3752 < \mu < 8.1592) \approx 0.95$?

Probabilities here are calculated for the random variable $\overline{X}$.
This expression has no random variable; it is not a meaningful expression.

7. Your 95% confidence interval is actually just the interval (7.3752, 8.1592). What do these numbers mean? Try to give a non-technical explanation of how to interpret this confidence interval.

We are 95% confident that this interval contains the pop. mean μ.
This means for 95% of samples, the interval we find actually includes μ. We can't tell whether this particular interval contains μ or not.

Definition. If $\overline{X}$ is the mean of a random sample of size n (with n large) from a population with mean μ and standard deviation σ, then the $100(1-\alpha)\%$ confidence interval for μ is $\boxed{\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$. Here $z_{\alpha/2}$ is the z-critical value: $P\left(Z > z_{\alpha/2}\right) = \alpha/2$ and can be found using the **R** command `qnorm(1` $-\ \alpha/2$`)`.

8. A random sample of 139 male house sparrows yields a sample mean blood plasma level (in pg/ml) of 209.46 and with a standard error of 16.62. Note that this is the standard error, not standard deviation: standard error is an estimate for the standard deviation of $\overline{X}$, usually $\frac{s}{\sqrt{n}}$. Use the standard error in place of $\frac{\sigma}{\sqrt{n}}$ in your confidence interval. Calculate 95% and 99% confidence intervals for the true mean plasma level of male house sparrows.

95% CI: $209.46 \pm 1.96(16.62)$ (176.8834, 242.0346)

99% CI: $209.46 \pm 2.5758(16.62)$ (166.6497, 252.2703)

Using a t-distribution:

95% CI: $209.46 \pm 1.977304(16.62)$ (176.5972, 242.3228)

99% CI: $209.46 \pm 2.611925(16.62)$ (166.0498, 252.8702)