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 sometimes other parameters). At this point, we have mostly thought 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 $c > 0$, $\lim\limits_{n\to\infty} P(|\hat{\Theta} - \theta| < c) = 1$, then we call $\hat{\theta}$ a **consistent estimator**.

Example. Suppose we know that a 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. For the sample below, the maximum is $\hat{x} = 5.0795233$.

0.8023426	1.1194331	1.5173202	1.5353951	2.1296237
2.5989737	3.4224997	3.7957449	4.8618741	5.0795233

It should make sense to use the maximum as an estimator for β. But is it an unbiased estimator? You'll need to use the CDF of the population: $F(x) = \dfrac{x}{\beta}$ if $0 < x < \beta$

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 $E(\hat{X})$. Is $\hat{X}$ an unbiased estimator for β?

d) Find a number c such that $E(c\hat{X}) = \beta$

Date: March 20, 2024.

The random sample above gives $2\overline{x} = 5.372546$ and $1.1\hat{x} = 5.587476$. (I'm using lowercase letters here because we have actual values for the sample statistics, so we're not thinking of them as random variables). We now have the values of two unbiased estimates for β. Which should we use? Is one of the estimates better 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**.

Challenge. Determine which of the estimators has a smaller variance: $2\overline{X}$ or $1.1\hat{X}$.

a) 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 $1.1\hat{X}$ by hand. You already know the expected value of $\hat{X}$, so it makes sense to use the formula $\text{Var}(\hat{X}) = E\left(\hat{X}^2\right) - \left[E(\hat{X})\right]^2$.

What we have been doing so far is called **point estimation**: 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 find a range of values that is likely to contain the population parameter.[1] This is called **interval estimation**.

2. Let Z be a standard normal random variable. Find a number z such that $\boxed{P(-z < Z < z) = 0.95}$
The R command `qnorm(p)` returns the $100p^{\text{th}}$ percentile of the standard normal distribution: this means $P(Z \leq \texttt{qnorm(p)}) = p$.

3. Substitute $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ into the boxed expression in problem 2, then isolate μ to fill in the blanks:

$$P(\overline{X} - \underline{\hspace{1.5cm}} < \mu < \overline{X} + \underline{\hspace{1.5cm}}) = 0.95$$

4. A random sample of size $n = 100$ is taken from a population with mean μ and standard deviation $\sigma = 2$. Suppose you find $\overline{x} = 7.7672$. Plug these numbers into your solution to the previous problem to find the **95% confidence interval** for the population mean μ.

[1] We'll have to be careful with the phrase "likely to contain the population parameter." Problems 5 and 6 deal with this.

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

6. The 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.

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 $\boxed{\overline{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}}$. 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(`$1-\alpha/2$`)`.

7. 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 $\frac{\sigma}{\sqrt{n}}$ to calculate confidence intervals. Calculate 95% and 99% confidence intervals for the true mean plasma level of male house sparrows.