Statistics ST2334 Topic 7: Inferences Concerning a Mean (part a)

Unbiased Estimators, Standard Error, Maximum Error of Estimate, Probability vs Confidence, Interval Estimation, Confidence Intervals.

2011/2012 Semester 2

Inference

- In the last chapter, we explored the sampling distribution of the mean. That is, how does the sample mean behave?
- Turns out we have a good understanding in most cases.
- However, one underlying assumption was that we knew the underlying population mean.
- The problem we are interested in however is usually the other way round.
- Given the sample mean, what can we say about the population mean?

Estimators

- An estimator is a rule, usually expressed as a formula, that tells us how to calculate an estimate based on information in the sample.
- We'll look at two types of estimators:
  - Point estimation: Based on sample data, a single number is calculated to estimate the population parameter.
    - The rule or formula that describes this calculation is called the point estimator.
    - the resulting number is called a point estimate.
  - Interval estimation: Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie.

Example: Point Estimator

- In order to estimate the average waiting time, μ, for a bus of a student attending ST2334, the lecturer asked 4 students their waiting time for a bus, the results are

$$X_1 = 6, X_2 = 1, X_3 = 4, X_4 = 9$$

- Use $\bar{X} = 5$ to estimate μ.
- $\bar{X}$ is the estimator, the computed value 5 is the estimate

Point Estimation of a Mean

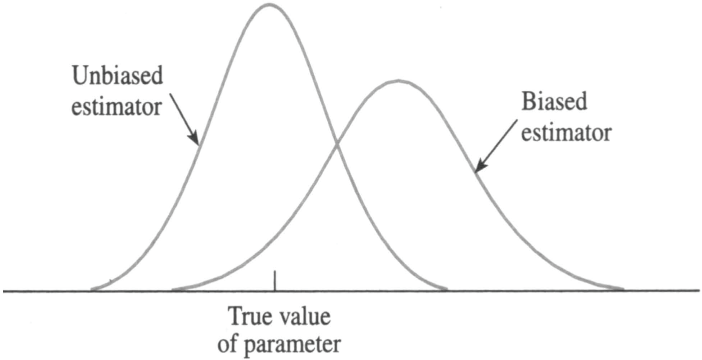
- Parameter of interest: μ, the population mean.
- Data: A random sample $X_1, X2, \ldots, X_n$.
- Estimator: $\bar{X}$
- How good is the estimator?

  Define the standard error (SE) of a point estimator to be the standard deviation of the underlying statistic.
  - So the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$
  - and if σ is unknown, we can estimate the SE as $s/\sqrt{n}$
  - because σ is usually unknown, the estimated SE is often simply called the SE.

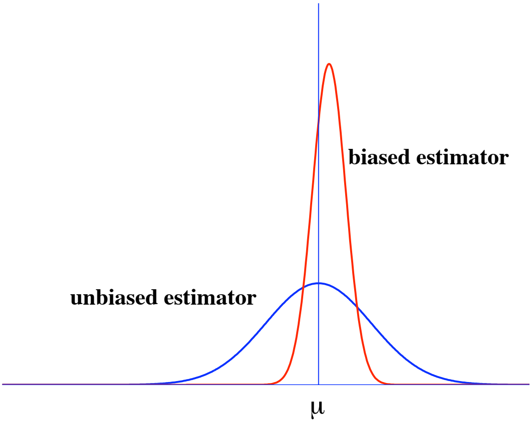
Unbiased Estimator

- One of the reasons we think $\bar{X}$ is a good estimator of μ is because $E(\bar{X}) = \mu$.
- That is, "on average", the estimator is right.
- In general, suppose θ is the parameter of interest. for example p, μ, σ
- let $\hat{\theta}$ be an estimator of θ, i.e. a random variable based on the sample.
- If $E(\hat{\theta}) = \theta$, we call $\hat{\theta}$ an unbiased estimator of θ.

Figures: Bias?



Figures: Bias?



- Unbiased estimator may not be better than biased estimator

Examples of Unbiased Estimators

If a bus arrives at the bus stop every θ (unknown) mins. The lecturer wants to estimate θ, so, this morning, he randomly selected 4 students and asked their waiting time for a bus. X_1, X_2, X_3, X_4 are obtained.

- Is $\bar{X}$ an unbiased estimator of θ? NO!

$$E(\bar{X}) = E(X) = \frac{\theta - 0}{2} = \theta/2 \neq \theta$$

- Find an unbiased estimator of θ. $2\bar{X}$ is one, since

$$E(2\bar{X}) = 2E(\bar{X}) = 2\theta/2 = \theta$$

- Is $2\bar{X}$ the only unbiased estimator of θ? NO! e.g: $2X_1$.

Unbiased Estimators?

$X_1, X_2, \ldots, X_n$ is a random sample from the same population with mean μ and variance σ^2.

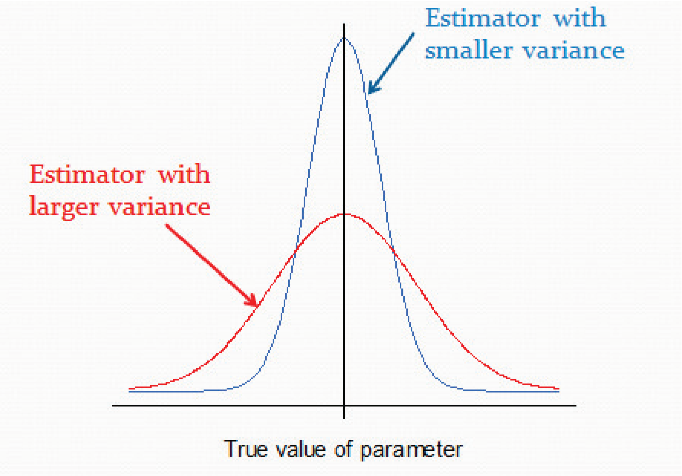
- Is $\bar{X}$ an unbiased estimator of μ?
- Is X_1 an unbiased estimator of μ?
- Is s^2 an unbiased estimator of σ^2?
- Is s an unbiased estimator of σ?
- Is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ an unbiased estimator of σ^2?
- Is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ an unbiased estimator of σ^2?

Comparing Unbiased Estimators

Usually we can find more than one possible unbiased estimators of a parameter θ. So which one is better?

- More efficient unbiased estimator.
- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ.
- If $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$ for all values of θ, and "$<$" is true for at least one value of θ
- $\hat{\theta}_1$ is said to be more efficient (better) estimator than $\hat{\theta}_2$.

Efficiency of Unbiased Estimators



Maximum Error of Estimate

- Usually $\bar{X} \neq \mu, \bar{X} - \mu$ measures the difference between the estimator and the true value of the parameter.
- Recall if the population is normal or n is large, then

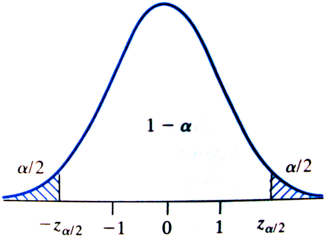
$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

either has standard normal or approximately follows a standard normal distribution.

Maximum Error of Estimate

- Therefore, with probability equal to $1 - \alpha$,

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$



Maximum Error of Estimate

- That is, with probability equal to $1-\alpha$,

$$\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

- Or with probability equal to $1-\alpha$, the error $|\bar{X}-\mu|$ is less than

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

  which is called Maximum Error of Estimate.
- Note that E increases with probability $1-\alpha$ and decreases with sample size n.

Example: Maximum Error of Estimate

- An investigator is interested in the amount of time internet users spend watching television per week. Based on historical experience, he assumes that the standard deviation $\sigma = 3.5$ hours.
- He proposes to select a random sample of $n = 50$ internet users, poll them, and take the sample mean to estimate the population mean μ.
- What can he assert with probability 0.99 about the maximum error of estimation?

  Since $n = 50$ (≥ 30) is a large value, $\sigma = 3.5$ and $z_{0.01/2} = 2.575$. With probability 0.99, the error is at most

$$E = 2.575(3.5/\sqrt{50}) \approx 1.27$$

Probability vs Confidence

- Now assume that the investigator collects the data and obtain $\bar{X} = 11.5$ hours.
- Can he still assert that with 99 % probability that the error is at most 1.27 hours?
- No! When talking about probability, always check where the randomness comes from. After the sample is taken, the error is at most 1.27 hours or not: there is no randomness!
- The probability is describing the method/estimator, not the result.
- To make this distinction, after $\bar{X}$ is obtained, we say instead that we conclude with 99% confidence, the error does not exceed 1.27 hours.

Determination of Sample Size

- What is the minimum n such that with probability (or with confidence afterwards) $1-\alpha$, the error is at most E? Since

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Solve for n, we have

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

- Example: what is the sample size n required such that the television investigator can assert with 99% probability that his estimation error is at most 0.5 hour?

$$n = \left(\frac{2.575 \cdot 3.5}{0.5} \right)^2 \approx 325$$

The Different Cases

- Recall we understood the sampling distribution of $\bar{X}$ for a variety of cases. Repeating the same arguments above, we have

Case	σ	n	Population	Statistic	E	sample size n needed for desired E and α
I	known	any	Normal	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$
II	known	large	any	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$
III	unknown	small	Normal	$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$	$t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{\alpha/2} \cdot s}{E}\right)^2$
IV	unknown	large	any	$Z = \frac{\bar{X}-\mu}{s/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot s}{E}\right)^2$

Back to television

- We are back to the case where the investigator polls $n = 50$ internet users.
- However, say we don't trust the historical assumption that $\sigma = 3.5$.
- Instead, we compute the sample standard deviation $s = 2.6$.
- With 99% confidence, what is E, the maximum error of our estimate?

$$E = z_{\alpha/2} \frac{s}{\sqrt{n}} = 2.575 \cdot \frac{2.6}{\sqrt{50}} \approx 0.947$$

Interval Estimation

- Since a point estimate is almost never right, one might be interested in asking for an interval where the parameter lies in.
- Interval estimator: a rule for calculating two numbers from the sample, say a and b, that create an interval (a, b) in which you are fairly certain the parameter of interest lies in.
- This "fairly certain" can be quantified by the degree of confidence also known as confidence level $(1 - \alpha)$, in the sense that

$$P(a < \mu < b) = 1 - \alpha$$

- (a, b) is called the $(1 - \alpha)$ confidence interval.

Case I: σ known, Data Normal

- We learnt that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and hence

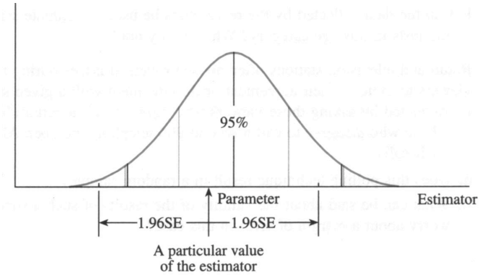
$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

Rearranging, we have

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Case I continued

- So
$$(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$
is a $(1-\alpha)$ confidence interval.
- It is often easier to write it as $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- Example: a 95% C.I. constructed from a sample



Example: CI

A computer company samples demand during lead time over 25 time periods:
235 374 309 499 253 421 361 514 462 369 394 439 348 344 330 261 374 302 466 535 386 316 296 332 334
It is known that the standard deviation of demand over lead time is 75 computers. Assuming the demands follows normal. We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels

- Form 95% confidence interval,

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 370.16 \pm 1.96 \frac{75}{\sqrt{25}} = 370.16 \pm 29.4$$

or can be written as $(340.76, 399.56)$.

Other Cases

- Notice that our $(1-\alpha)$ CI can be written as

$$\bar{X} \pm E$$

- This is not a coincidence: recall there is $(1-\alpha)$ probability/confidence that the error $|\bar{X}-\mu|$ is within E.
- For the other cases, based on our understanding of the sampling distribution of $\bar{X}$, we can construct our confidence intervals for the different cases $\bar{X} \pm E$, based on the conditions given (see slide 19).

Example: Which Case?

- The following data set collects n=41 randomly sampled waiting times of students from ST2334 to receive reply for their email from a survey in the day time.

2.50	23.28	19.34	4.74	7.03	21.85	2.72	10.64
17.73	21.55	9.71	30.24	0.37	31.26	35.24	13.10
7.81	16.69	66.54	1.88	14.14	46.59	28.17	7.92
0.06	9.32	0.03	10.75	6.97	56.86	2.89	112.77
7.67	30.16	0.33	0.44	3.77	25.07	7.05	11.93
0.08							

- Question: construct 98% C.I. for the mean waiting time of students.

Example: Which Case?

- σ unknown, n large $\Rightarrow$ Case IV.
- our 98% CI is

$$\begin{aligned}
\bar{X} \pm E &= \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \\
&= 17.736 \pm 2.33 \frac{21.7}{\sqrt{41}} \\
&= (9.84, 25.63)
\end{aligned}$$

Interpreting Confidence Intervals

- We have argued that $\bar{X} \pm E$ has probability $(1-\alpha)$ of containing μ.
- Note that this is a probability statement about the interval estimator.
- Once an interval is established, μ is either in it or not. There is no more randomness.
- That is, once we collect a sample and compute the actual confidence interval (numbers), we can only talk about our interval having $(1-\alpha)$ confidence. Not probability!

What does this confidence really mean?

- It is a reminder that the randomness is in the procedure/method, and not the parameter.
- If we take another sample, and go through this construction again, we get a different CI.
- If we repeat this many times, about $(1 - \alpha)$ of the many CI's that we get will contain the true parameter.

Example: Many confidence intervals