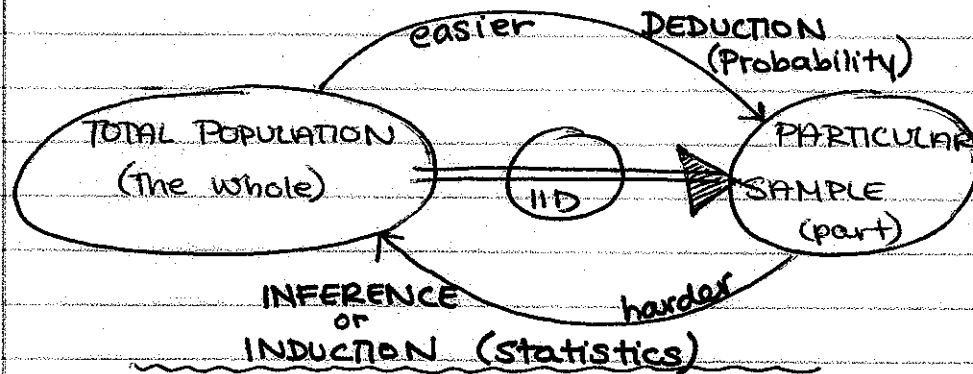


STATISTICAL INFERENCE

difference between probability and statistics

- **DEDUCTIVE REASONING (PROBABILITY)** → making a conclusion about a sample by looking at the total population (what we have been doing).
- **INDUCTIVE REASONING: (STATISTICS)** → using the results from a sample to make a conclusion about the total population.



STATISTICAL INFERENCE: Reasoning from the factual to the counterfactual

- We know how the sample came out $\& \uparrow$ we want to reason back to the population
- The same basic model is used for probability and statistical cases; only difference is in what is known $\& \uparrow$ what is unknown.

EXAMPLE

Intertidal crabs live both in the water and in the air, so they must change several bodily functions, including their body T° to survive.

EXPERIMENT → 25 intertidal crabs were left in the air whose T° is 24.3°C . They were allowed time to equilibrate their body T° 's, and then their body T° 's were measured.

THEORY → The crabs should equilibrate their body T° 's so that they match the air T° of 24.3°C

* T° = temperature

RESULTS FOR 25 INTERTIDAL CRABS:

MEAN (\bar{y}) = 25.0°C

SD (s) = 1.34°C

DIFFERENCE = (actual - theoretical) = $25^{\circ}\text{C} - 24.3^{\circ}\text{C} = 0.7^{\circ}\text{C}$

Is this difference big in practical terms?

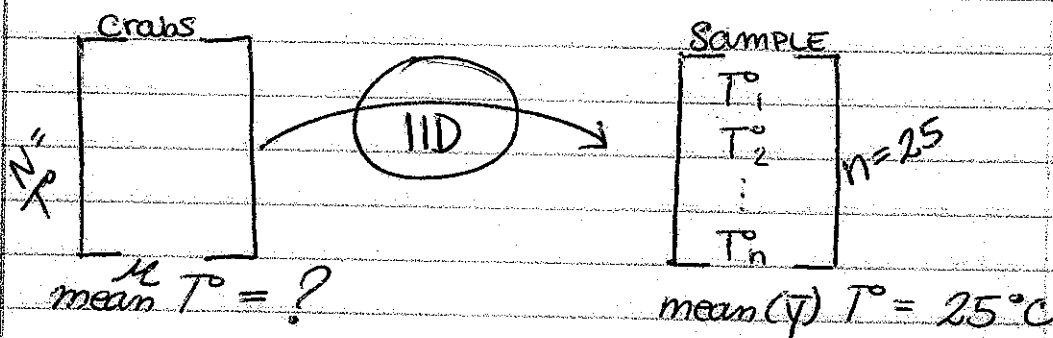
- Why isn't the mean 24.3°C - Can the crabs' body T° change by 0.7°C without harming them?- Is $7/10^{\circ}\text{C}$ a big difference for crabs?Is the difference big in statistical terms?

- Is the difference just a case of unlucky random sampling?

- Is the data evidence running in support of or against the theory?

Answering the question in practical terms is very important but now we focus on the question from a statistical point of view.

POPULATION

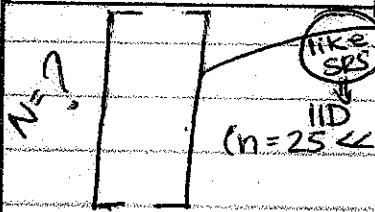


We know how the sample came out and we want to reason back to the population.

- Use same basic model w/ different knowns \rightarrow and unknowns

POPULATION

Sample represents all intertidal crabs that are similar to those sampled.



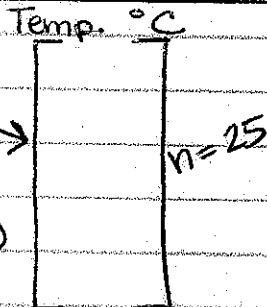
$\mu = ?$
 $\sigma = ?$

- σ of population should be smaller than s of sample
 $n \uparrow, SE \downarrow$
histogram = ?

Because the sample is related to the population, these should be similar to the sample

SAMPLE

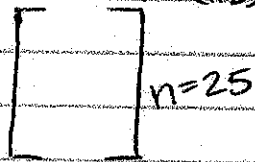
The temperatures of the observed crabs



$\bar{y} = 25^\circ\text{C}$
 $s = 1.34^\circ\text{C}$

To make an imaginary data set, we want to hypothetically repeat the sample process M times to get M sample means

HYPOTHETICAL (IID)

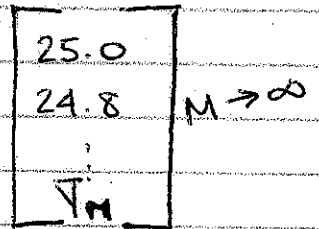


$\bar{y} = ?$ (ex: 24.8°C)

(more hypothetical samples)

IMAGINARY DATA SET

All the possible \bar{y} 's.



long run mean:

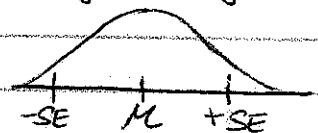
$E_{IID}(\bar{y}) = \mu$

long run SD:

$SE_{IID}(\bar{y}) = \frac{\sigma}{n}$

long run histogram:

use CLT if n is large enough.



WE DON'T KNOW ANY OF THE IMAGINARY DATA SET RESULTS

HOW TO FIND IMAGINARY DATA

- We do not know the SE because we do not know σ . However, because the sample is related to the population, σ should be similar to S .

$$\text{Estimated SE of } \bar{y} = \hat{SE}_{11D}(\bar{y}) = \frac{S}{\sqrt{n}}$$

(Rather than $SE_{11D}(\bar{y}) = \frac{\sigma}{\sqrt{n}}$)

$$\hat{SE}_{11D}(\bar{y}) = \frac{S}{\sqrt{n}} = \frac{1.34}{\sqrt{25}} = 0.27$$

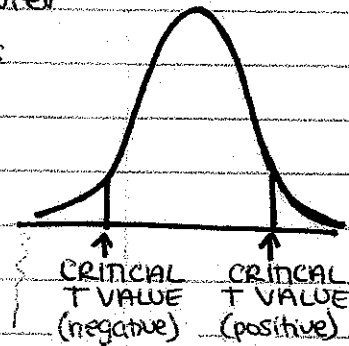
Because you cheat in finding the SE, maybe the normal curve is not your best choice to approximate to.

- LONG RUN HISTOGRAM OF \bar{y}

- must account for uncertainty in σ

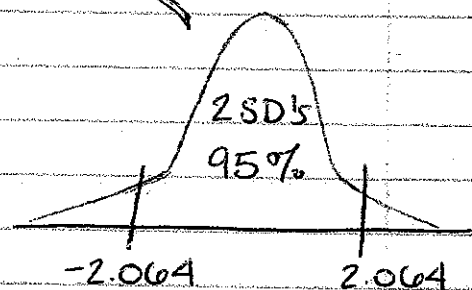
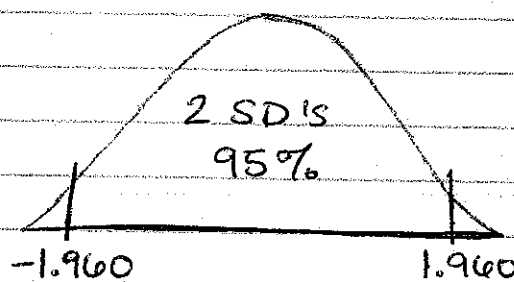
THE T CURVE

- still symmetrical, but the uncertainty would be bigger \rightarrow make the tails heavier
- different T curve for different sample sizes (same way the normal curve works \rightarrow relate to one standard T curve)
- as $n \uparrow$, T curve should look more like normal curve because SE goes down.



- price must be paid when n is small

- T curve is used on $(n-1)$ degrees of freedom and has a t Table just like normal curve

NORMAL CURVE VS. T_{n-1} CURVE

CONFIDENCE INTERVAL (CI) → What is the probability that μ will be found within the \hat{SE} of (\bar{y})

- Use \bar{y} as an estimate of μ and see whether the theoretical μ is within the 95% Confidence Interval.

95% CI for μ μ will differ from \bar{y} by no more than $(2SD's) \times (\hat{SE})$ with a probability of 95%.

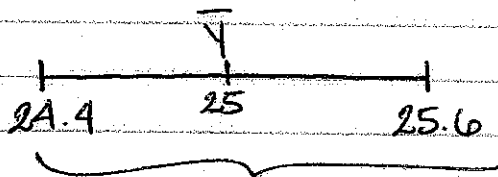
$$\bar{y} \pm (t_{n-1}^{0.95}) (\hat{SE}_{\bar{y}})$$

Place on the T curve with $(n-1)$ degrees of freedom where 95% of the area is in the middle = 2SDs

$$= \bar{y} \pm (2.064) \left(\frac{s}{\sqrt{n}}\right)$$

- People use a 95% CI (2SDs) because it became standard as high confidence a long time ago
 - conventional because people have 5 fingers on one hand
 - 95% is close to 100% but differs by 5 fingers
- In any case, $CI < 100\%$ because $CI = 100\%$ means that mean (μ) is between $-\infty$ & $+\infty$
 - giving a high CI would give the μ as too broad
 - giving a lower CI would give a narrower breadth for the μ but the CI may not contain the μ .

$$\begin{aligned} CI \text{ of } \mu \text{ for INTERTIDAL CRABS} &= \bar{y} \pm 2.064 \left(\frac{s}{\sqrt{n}}\right) \\ &= 25 \pm 2.064(0.27) = 24.4^\circ\text{C} \\ &\quad \updownarrow \\ &\quad 25.6^\circ\text{C} \end{aligned}$$



μ should be somewhere in here w/ 95% confidence

02/10/09

INFERENTIAL SUMMARY

Unknown quantity of interest in the population.

μ = mean T° @ which intertidal crabs of a specific species would equilibrate when air $T^\circ = 24.3^\circ\text{C}$

estimate of μ

$$\bar{y} = 25.0^\circ\text{C}$$

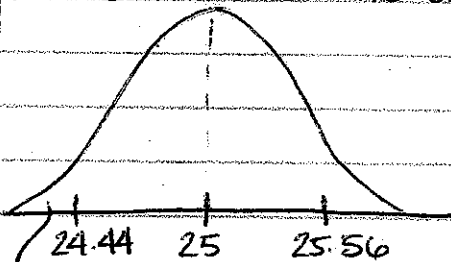
give or take for \bar{y} as an estimate for μ

$$\hat{SE}_{\text{HD}}(\bar{y}) = \frac{s}{\sqrt{n}} = 0.27^\circ\text{C}$$

95% CI for μ

$$\bar{y} \pm (t_{n-1}^{0.95}) \left(\frac{s}{\sqrt{n}} \right) = 24.4^\circ\text{C} \text{ } 25.6^\circ\text{C}$$

CONCLUSION \rightarrow The data does not support the theory



$24.3^\circ\text{C} =$ theoretical value for $\mu = \mu_0$

Because μ_0 is not in the 95% confidence interval, the data does not support the theory @ the 95% confidence level

Difference of 0.7°C is statistically significant (large in statistical terms)

$$0.7^\circ\text{C} = \bar{y} - \mu_0$$

SUMMARY OF NEW FORMULAS

$$\hat{SE}_{\text{HD}}(\bar{y}) = \frac{s}{\sqrt{n}} \quad (s = \text{sample standard deviation})$$

$$95\% \text{ CI} = \bar{y} \pm (t_{n-1}^{0.95}) (\hat{SE}_{\text{HD}}(\bar{y}))$$