Maya Sertic

2/12/09

STATSIG & PRACTSIG

To see whether a single sample ($Y_1 - Y_n$) from a population w/ mean μ supports the theory that $\mu = \mu_0$ (where μ_0 = theoretical value), build a 95% confidence interval for μ using the T machinery & see if μ_0 is in the interval.

- "Is the difference real?" asks if the difference is statistically significant (statsig). The difference is statististically significant when μ_0 is not in 95% CI. If μ_0 not in 99% CI, difference is highly statsig.
- STATSIG = hard to attribute to unlucky random sampling

Before we ask about statistical significance, we should always first ask if the difference is PRACTICALLY SIGNIFICANT (practsig)

- We may not always know the answer to this question ourselves, so we must talk to somebody who would know about the practical significance of the difference.

If the difference is not practsig, then there is no point in looking into whether the difference is statsig.

Last time we looked at a population (intertidal crab body T°) where the data was quantitative continuous. How can we use the same strategies when the population is binary?

→ EXAMPLE → lab rats choosing which way to go for food

Maya Sertic

02/12/09

Example:

12 lab rats were tested in terms of decision making based on smell.

- Put into a maze where turning left = turning toward food & turning right = turning away from food.

THEORY → rats would choose at random which direction to go in → 50% will go left & 50% will go right.

ACTUALLY → 10 rats turned left (83%)
2 rats turned right

Is this difference PRACTSIG?

- you would ask a biologist, & he would likely say yes → huge difference.

STATSIG? → Start w/ basic model

POPULATION

all relevant labrats (ask biologist)

$N = ?$ (big) $\begin{bmatrix} 1\text{'s } \& \\ 0\text{'s} \end{bmatrix}$

like SRS ≠ IID, $n << N$

mean $\mu = p = ?$

SD $\sigma = ?$

histogram ?

like IID

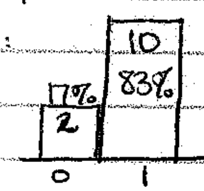
SAMPLE

Choice of the observed labrats (Left = 1, Right = 0)

(actual) $\begin{bmatrix} 1\text{'s} \\ \& \\ 0\text{'s} \end{bmatrix}$ $n = 12$

mean $\bar{Y} = \hat{p} = \frac{10}{12} = 83\%$

histogram:



(hypothetical)

$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{12} \end{bmatrix}$ $n = 12$

$\hat{p} = ?$ (ex = 75%)

IMAGINARY DATA SET

all possible $\hat{p}$ values

$\hat{P} = \bar{Y}$

$\begin{bmatrix} 83\% \\ \vdots \\ \hat{P}_M \end{bmatrix}$ $M \to \infty$

long run mean = $E_{IID}(\hat{P}) = P$ $\boxed{1}$

long run SD = $\hat{SE}_{IID}(\hat{P}) = 11\%$ $\boxed{2}$

long run histogram $\boxed{3}$

Maya Sertic

2/12/2009

Evaluating Long Run Data

$\boxed{1}$ EV of $\hat{p} = E_{IID}(\hat{p}) = E_{IID}(\bar{Y}) = \mu = p$

$$E_{IID}(\hat{p}) = p$$

$\boxed{2}$ $SE(\hat{p}) = SE_{IID}(\hat{p}) = SE_{IID}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$

- If a population has only 2 values in it, pop. SD $= \sigma =$ (larger value) − (smaller value) $\sqrt{\left(\text{proportion of larger values}\right)\left(\text{proportion of smaller values}\right)}$
  - here, larger value = 1 &
    smaller value = 0

w/ binary population, $SD = \sigma = \sqrt{p \cdot (1-p)}$

- $SE_{IID}(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

In a binary population, $SE_{IID}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

Because we do not know what p is, we cannot find the SE. Instead, we can use our cheating method and find $\hat{SE}$.

$$\hat{SE}_{IID}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{(0.837)(0.17)}{12}} = 0.108 \approx 11\%$$

CI for continuous outcomes $= \bar{Y} \pm (t_{n-1}^{0.95})(\hat{SE}_{IID}(\bar{Y}))$

We need a CI for a binary outcome.

If n is large, then the long run histogram of $\hat{p}$ is a normal curve. (CLT) When n is small, you do not use the normal curve, but the curve for this is too technical for the scope of this class

- APPROXIMATE METHOD : pretend n is big enough to use the CLT → histogram will be a normal curve.

$$CI = \hat{p} \pm (2SDs)(\hat{SE}_{IID}(\hat{p})) = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

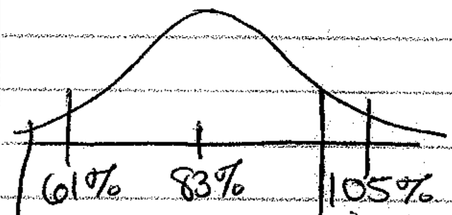
In our lab rat case:

$$95\%\ CI = 83\% \pm 1.96(11\%) = 61\% \text{ \& } 105\%$$

Maya Sertic

4/5
02/12/09

lab rat case:



Truncate @ 100%

50% is not in 95% CI
difference between 83% and theory of 50% is statsig and practsig.

INFERENTIAL SUMMARY

quantity of interest for unknown population	p = population % of rats similar to these that would turn left (toward food)
estimate of p	$\hat{p} = 83\%$
give or take for $\hat{p}$ as estimate for p	$\hat{SE}_{IID}(\hat{p}) = \hat{SE}_{IID}(\bar{Y}) = \frac{\hat{\sigma}}{n} = 11\%$
approximate 95% CI for p.	$\hat{p} \pm 1.96(\hat{SE}_{IID}(\hat{p})) = 61\% \,\&\, 100\%$

There are other ways to do statistical inference, but confidence interval is the best way

- HYPOTHESIS TESTS
- SIGNIFICANCE TESTS

} talk about these just to learn language of stats. & be able to work w/ scientists.

To talk about these, go back to crab example.

Maya Sertic

5/5
02/12/09

HYPOTHESIS TEST

$\boxed{1}$ NULL HYPOTHESIS $\mu = 24.3 = \mu$ (theory correct)

$\boxed{2}$ ALTENATIVE HYPOTHESIS $\mu \neq 24.3\,^{\circ}C$ (theory wrong)

⟶ Process is to choose one hypothesis and decide how probable the data is if that hypothesis is correct.

- Assume null hypothesis is correct because is uses less math (easier)

$\boxed{1}$ → difference between $\mu_0 = 24.3\,^{\circ}C$ and $\bar{Y} = 25\,^{\circ}C$ is due to unlucky random sampling. (logical possibility)

- This is the statistical version of the "Proof by Contradiction" in calculus, only for probabilities.

See if the discrepancy between

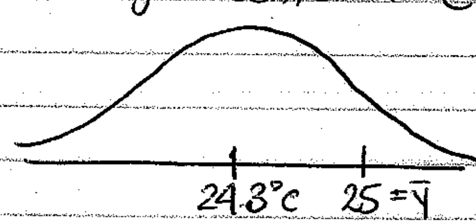
(how data came out *) vs. (how data should have come out **)

is large.

- If yes, favor the alternative hypothesis (reject null).
- If not, favor the null (fail to reject null)
  - If the null is true, nothing can be said about the sample

For total population, if null is true, $\mu = 24.3 = \mu_0$

- In Imaginary data set, EV of $\bar{Y} = \mu_0 = 24.3$
- Long Run Histogram no longer centered @ $\mu = ?$, but at $\mu = 24.3$



TO BE CONTINUED...