

Statistics

... is all about making inferences concerning a population through the use of sample data.

... must therefore worry about exposure to *sampling error*, i.e., bad luck in the sampling process which leads to a somewhat-misrepresentative sample.

... deals with this problem by looking at the end results of statistical procedures as random variables, and using the tools of probability to study these random variables.

Estimation

Estimating a population mean, using simple random sampling:

estimate		margin of error (at 95%-confidence)
\bar{x}	\pm	$(\sim 2) \cdot \frac{s}{\sqrt{n}}$

Estimating a proportion, using simple random sampling:

	estimate		margin of error (at 95%-confidence)
precisely	\hat{p}	\pm	$(\sim 2) \cdot \frac{\sqrt{\hat{p} \cdot (1-\hat{p})}}{\sqrt{n}}$
conservatively	\hat{p}	\pm	<i>(less than)</i> $\frac{1}{\sqrt{n}}$

The estimate is typically used for decision analysis, and the margin of error for risk/sensitivity analysis.

More generally, when estimating (or predicting) anything:

estimate		margin of error (at 95%-confidence)
<i>(the estimate)</i>	\pm	$(\sim 2) \cdot$ <i>(the standard error of the estimate)</i>

where the ***standard error of the estimate*** is one standard-deviation's-worth of potential error in the estimate due to uncertainties inherent in the estimation procedure.

Setting the Sample Size (when using simple random sampling)

When estimating a **mean** or **proportion** using simple random sampling, *with a preliminary study at hand*:

$$(\text{needed sample size}) = \left(\frac{\text{current margin of error}}{\text{target margin of error}} \right)^2 \cdot (\text{current sample size})$$

The preliminary study at hand could be a past study, or a pilot study.

When estimating a **proportion** using simple random sampling:

The actual margin of error (at the 95%-confidence level) will always be somewhat less than $1/\sqrt{n}$, so setting

$$\text{sample size} = \frac{1}{(\text{target margin of error})^2}$$

will always suffice.

Hypothesis Testing

We begin with a statement – the null hypothesis – “on trial.” At the end of the trial, we will find that the evidence at hand either contradicts the statement to some extent (i.e., the evidence supports a finding of “guilty”), or doesn’t really contradict the statement (i.e., the evidence doesn’t support a finding of “guilty,” so we find it “not (shown to be) guilty”).

Think of a hypothetical world in which the statement on trial is true. (If there’s more than one such world, choose the one which most closely fits the observed data.)

The *significance level* of the data (with respect to the statement on trial) is

Prob (we’d see data at least as contradictory to the statement as is the data at hand | the study that yielded the data at hand were to be conducted in the hypothetical world where the statement is true)

We interpret the significance level of the data using this “translation” table:

If the numeric significance level of the data is	then the data, all by itself, makes us	and the data supports the alternative
above 20%	not at all suspicious	not at all
between 10% and 20%	a little bit suspicious	a little bit
between 5% and 10%	moderately suspicious	moderately
between 2% and 5%	very suspicious	strongly
between 1% and 2%	extremely suspicious	very strongly
below 1%	overwhelmingly suspicious	overwhelmingly

We never conclude that the evidence *supports* the statement on trial (i.e., the statement is never found “innocent”). Therefore, if our ultimate goal is to see if evidence supports a statement, we must put the opposite statement on trial, and see if the evidence contradicts that opposite statement.