Inference in practice

BPS chapter 16

Objectives (BPS chapter 16)

Inference in practice

- Where did the data come from?
- Cautions about *z* procedures
- Cautions about confidence intervals
- Cautions about significance tests
- The power of a test
- Type I and II errors

Where did the data come from?

- When you use statistical inference, you are acting as if your data are a probability sample or come from a randomized experiment.

- **Statistical confidence intervals and hypothesis tests cannot remedy basic flaws in producing the data**, such as voluntary response samples or uncontrolled experiments.

Caution about z procedures

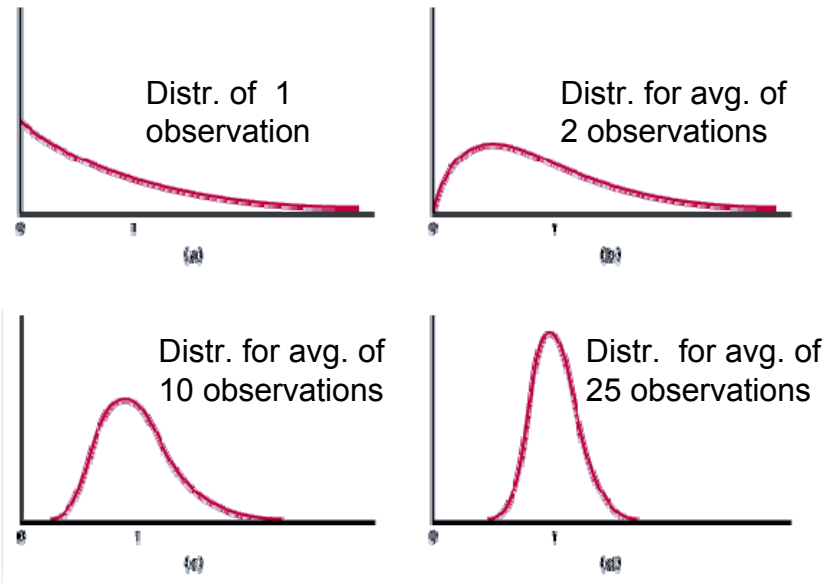
Requirements

- **The data must be an SRS,** simple random sample, of the population. More complex sampling designs require more complex inference methods.

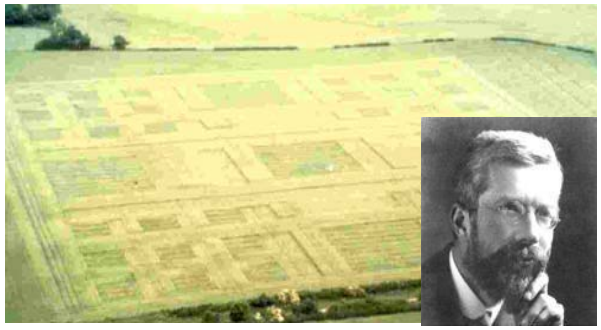
- **The sampling distribution must be approximately normal.** This is not true in all instances.

- **We must know** σ, the population standard deviation.
  This is often an unrealistic requisite.
  We'll see what can be done when σ *is unknown in the next chapter.*

- We cannot use the *z* procedure if the population is not normally distributed and the sample size is too small because the central limit theorem will not work and the sampling distribution will not be approximately normal.



- Poorly designed studies often produce useless results (e.g., agricultural studies before Fisher). Nothing can overcome a poor design.



- Outliers influence averages and therefore your conclusions as well.

Cautions about confidence intervals

The margin of error does not cover all errors: The margin of error in a confidence interval covers only ***random sampling*** error.

Undercoverage, nonresponse, or other forms of bias are often more serious than random sampling error (e.g., our elections polls). The margin of error does not take these into account at all.

Cautions about significance tests

How small a *P*-value is convincing evidence against H_0?

Factors often considered in choosing the significance level α:

- What are the consequences of rejecting the null hypothesis (e.g., global warming, convicting a person for life with DNA evidence)?
- Are you conducting a preliminary study? If so, you may want a larger alpha so that you will be less likely to miss an interesting result.

Some conventions:

- We typically use the standards of our field of work.
- There are no “sharp” cutoffs: e.g., 4.9% versus 5.1 %.
- It is the order of magnitude of the P-value that matters (“somewhat significant,” “significant,” or “very significant”).

Practical significance

Statistical significance only says whether the effect observed is likely to be due to chance alone because of random sampling.

Statistical significance may not be practically important. That's because statistical significance doesn't tell you about the **magnitude** of the effect, only that there is one.

An effect could be small enough to be irrelevant. And with a large enough sample size, a test of significance can detect even very small differences between two sets of data, as long as it is real.

- Example: Drug to lower temperature, found to reproducibly lower a patient's temperature by 0.4° Celsius (*P*-value < 0.01). But clinical benefits of temperature reduction, found to appear for 1° decrease or more.

Sample size affects statistical significance

- Because **large random samples** have small chance variation, very small population effects can be highly significant if the sample is large.

- Because **small random samples** have a lot of chance variation, even large population effects can fail to be significant if the sample is small.

Interpreting effect size: It's all about context

There is no consensus on how big an effect has to be in order to be considered meaningful. In some cases, effects that may appear to be trivial can in reality be very important.

- Example: Improving the format of a computerized test reduces the average response time by about 2 seconds. Although this effect is small, it is important since this is done millions of times a year. The *cumulative* time savings of using the better format is gigantic.

Always think about the context. Try to plot your results, and compare them with a baseline or results from similar studies.

More cautions…

Confidence intervals vs. hypothesis tests

- It's a good idea to give a confidence interval for the parameter in which you are interested. A confidence interval actually estimates the size of an effect rather than simply asking if it is too large to reasonably occur by chance alone.

Beware of multiple analyses

- Running one test and reaching the 5% level of significance is reasonably good evidence that you have found something. Running 20 tests and reaching that level only once is not.
  - *A single 95% confidence interval has probability 0.95 of capturing the true parameter each time you use it.*
  - *The probability that all of 20 confidence intervals will capture their parameters is much less than 95%: it is $(0.95)^{20}$ = only 0.358.*
  - *If you think that multiple tests or intervals may have discovered an important effect, you need to gather new data to do inference about that specific effect.*

The power of a test

The **power** of a test of hypothesis with fixed significance level α is the probability that the test will reject the null hypothesis when the alternative is true.

In other words, power is the probability that the data gathered in an experiment will be sufficient to reject a wrong null hypothesis.

Knowing the power of your test is important:

- When designing your experiment: To select a sample size large enough to detect an effect of a magnitude you think is meaningful.
- When a test found no significance: Check that your test would have had enough power to detect an effect of a magnitude you think is meaningful.

How large a sample size do I need?

How large a sample do we need for a z test at the 5% significance level to have a power 90% against various effect sizes?

Effect size	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Sample size	857	215	96	54	35	24	18	14	11	9

When calculating the effect size, think about how large an effect in the population is important in practice.

$$\text{effect size} = \frac{\text{true mean response - hypothesized response}}{\text{standard deviation of response}}$$

How large a sample size do I need?

In general:

- If you want a smaller significance level (α) or a higher power (1 - β), you need a larger sample.

- A two-sided alternative hypothesis always requires a larger sample than a one-sided alternative.

- Detecting a small effect requires a larger sample than detecting a larger effect.

Test of hypothesis at significance level α 5%:

H_0: $\mu = 0$ versus H_a: $\mu > 0$

Can an exercise program increase bone density? From previous studies, we assume that $\sigma = 2$ for the percent change in bone density and would consider a percent increase of 1 medically important .
Is 25 subjects a large enough sample for this project?

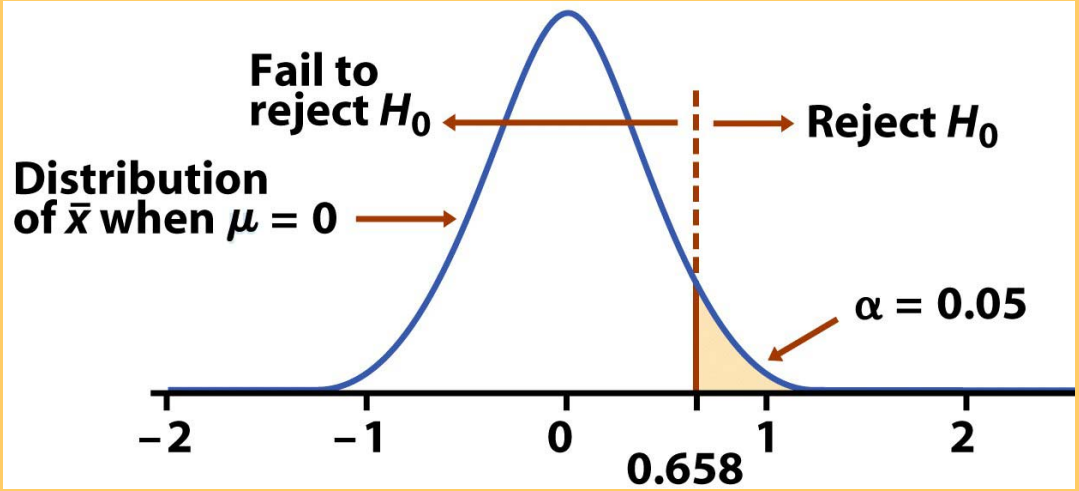
A significance level of 5% implies a lower tail of 95% and $z = 1.645$. Thus:

$$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$$

$$\bar{x} = \mu + z^{*}(\sigma / \sqrt{n})$$

$$\bar{x} = 0 + 1.645^{*}(2 / \sqrt{25})$$

$$\bar{x} = 0.658$$



All sample averages larger than 0.658 will result in rejecting the null hypothesis.

What if the null hypothesis is wrong and the true population mean is 1?

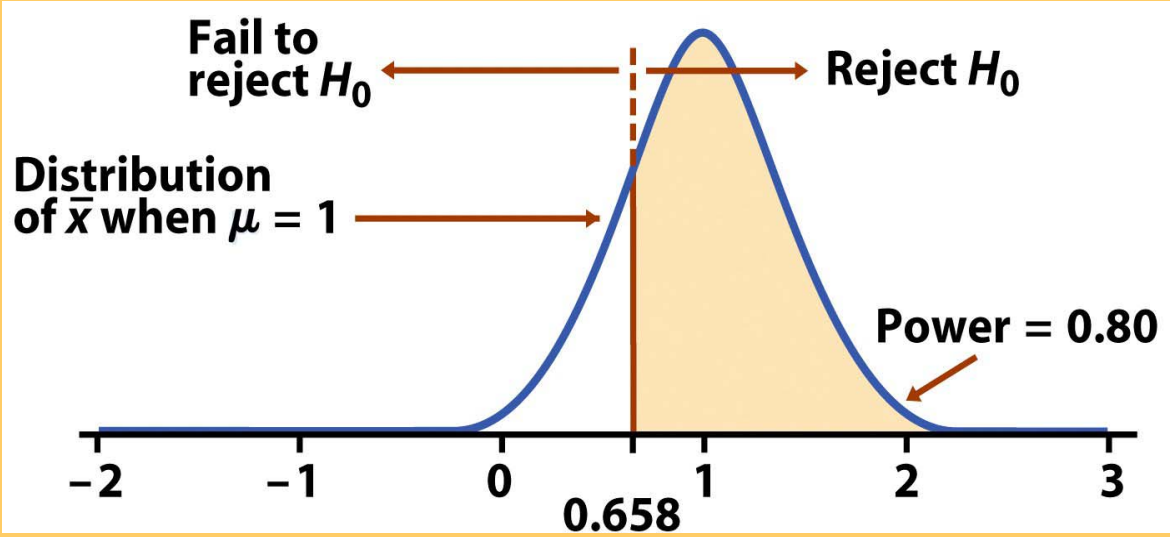
The **power against the alternative** μ = 1% is the probability that H_0 will be rejected when in fact μ = 1%.

$$= P(\bar{x} \geq 0.658 \text{ when } \mu = 1)$$

$$= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.658 - 1}{2/\sqrt{25}}\right)$$

$$= P(z > -0.855) = 0.80$$

We expect that a sample size of 25 would yield a power of 80%.



A test power of 80% or more is considered good statistical practice.

Type I and II errors

- A **Type I error** is made when we reject the null hypothesis and the null hypothesis is actually true (incorrectly reject a true H_0).

  The probability of making a Type I error is the significance level α.

- A **Type II error** is made when we fail to reject the null hypothesis and the null hypothesis is false (incorrectly keep a false H_0).

  The probability of making a Type II error is labeled β.

  The power of a test is $1 - \beta$.

Type I and II errors—court of law

H_0: The person on trial is not a thief.

(In the U.S., people are considered innocent unless proven otherwise.)

H_a: The person on trial is a thief.

(The police believe this person is the main suspect.)

- A **Type I error** is made if a jury convicts a truly innocent person.

  (They reject the null hypothesis even though the null hypothesis is actually true.)

- A **Type II error** is made if a truly guilty person is set free.

  (The jury fails to reject the null hypothesis even though the null hypothesis is false.)

Running a test of significance is a balancing act between the chance α of making a **Type I error** and the chance β of making a **Type II error.** Reducing α reduces the power of a test and thus increases β.

	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

It might be tempting to emphasize greater power (the more the better).

- However, with "too much power" trivial effects become highly significant.
- A Type II error is not definitive since a failure to reject the null hypothesis does not imply that the null hypothesis is wrong.