Statistical Inference

"All models are wrong but some are useful" George Box

Statistical models are used to describe the underlying processes that generate sample data. Depending on the type of data, a potential model is selected from a family of models and the data is used to estimate model parameters. The fitted model is then compared to alternative models and the most appropriate model is selected. Once a statistical model has been chosen, the estimated model can be used to make inferences from the data: testing hypotheses, creating predicted values, measures of confidence. The estimated model becomes the lens through which we interpret the data.

Estimators and Estimates

Recall, Statistical inference uses quantities computed from the observations in the sample.
A **point estimator** is defined to be a function $h\left(X_1, X_2, ..., X_n\right)$ of the random sample $X_1, X_2, ..., X_n$ that can used to estimate a parameter of the underlying population
An **estimate** refers to the actual value computed from the observed data set, $x_1, x_2, ..., x_n$.
e.g. the point estimator for the population mean μ is the sample mean $\overline{X}$:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

An estimator is itself a random variable with an associated probability distribution.

Hypothesis Testing

Recall, a statistical hypothesis is a statement about the values of the parameters of a probability distribution. For example, suppose that we think the average pH of a particular solution is 5. We may express this statement in a formal manner as

$$\begin{aligned} H_0 &: \mu = 5 \\ H_A &: \mu \neq 5 \end{aligned}$$

The statement $H_0 : \mu = 5$ is called the **null hypothesis**
The statement $H_A : \mu \neq 5$ is called the **alternative hypothesis**

It is important to remember that hypotheses are always statements about the population or distribution under study, not statements about the sample.

Note that any scientific investigation is subject to errors and uncertainties, and the results of a hypothesis test **do not give an absolute** answer to the question posed, instead we assess the results of a hypothesis test in terms of probabilities. There is never an absolute certainty in the answer but it is possible to calculate the probability of making an error.

Errors

When performing a hypothesis test, there are two ways of being wrong:

1. **type I error** reject the null hypothesis when it is in fact true
2. **type II error** fail to reject the null hypothesis when it is in fact false

$$\begin{aligned} P\,(\text{type I error}) &= P\{\text{reject } H_0 |\ H_0 \text{ is true}\} \\ P\,(\text{type II error}) &= P\{\text{fail to reject } H_0 |\ H_0 \text{ is false}\} \end{aligned}$$

The probability of a type I error for a given experiment is given by the **p-value**, which can be calculated from the experimental results.
The probability of a type II error is denoted by β and is much more difficult to calculate.

It is essential, before performing any experimental measurements, to decide on the probability of error that would be acceptable when making the decision. Before starting the experiment, the **significance level** α must be stated. The significance level α is the maximum probability of a type I error that would be acceptable. A decision is made based on the relative values of α and the calculated p-value.

$$\begin{aligned} \text{If p-value } &\leq \alpha \text{ reject the null hypothesis, } H_0, \text{ at a significance level of } \alpha \\ \text{If p-value } &> \alpha \text{ fail to reject } H_0 \end{aligned}$$

p-values can be difficult to calculate but can be checked using many software packages such as R or Minitab.

Inference for a Difference in Means - Variances Unknown (the t-test revisited) Suppose that X is a random variable with unknown mean μ **and unknown variance** σ^2. Suppose that we wish to test whether two means μ_1 and μ_2 are equal where the population variances σ_1^2 and σ_2^2 are unknown.

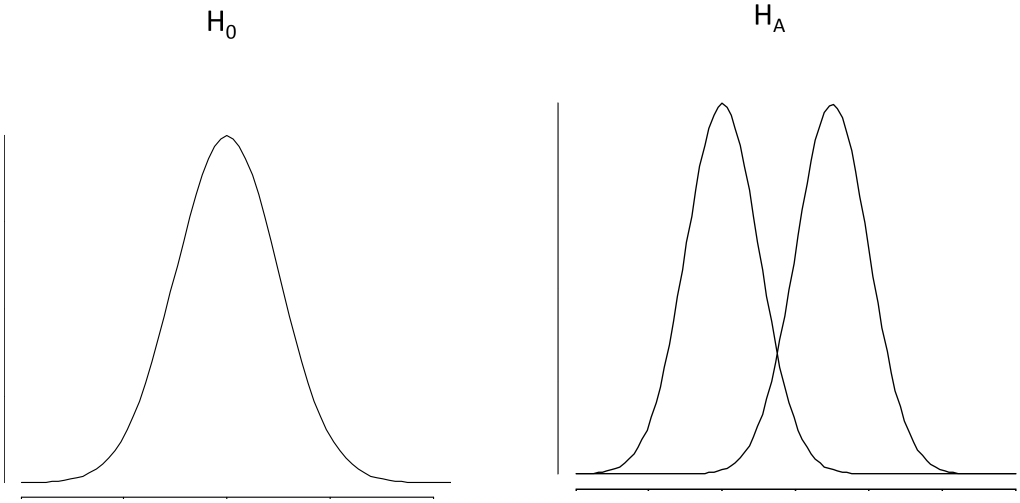


Fig. 1

The hypothesis are:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

The test procedure depends on whether $\sigma_1^2 = \sigma_2^2$. If this is the case, then a random sample of n_1 observations from population 1 and a random sample of n_2 observations from population 2 is taken and a **pooled** estimate of the variance is obtained using the formula

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

where s_1^2 and s_2^2 are the individual sample variances. The test statistic is:

$$t_{stat} = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The test statistic will follow a t distribution with $n_1 + n_2 - 2$ degrees of freedom (Fig. 2)

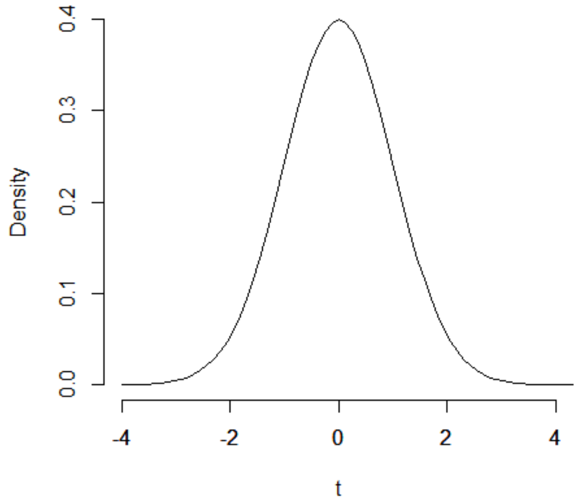


Fig. 2

We need to check whether $|t_{stat}| > t_{\alpha/2,n_1+n_2-2}$ where $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t distribution with $n_1 + n_2 - 2$ degrees of freedom. If

$$\begin{aligned} |t_{stat}| &> t_{\alpha/2,n_1+n_2-2} \ : \ \text{reject } H_0 \\ |t_{stat}| &< t_{\alpha/2,n_1+n_2-2} \ : \ \text{fail to reject } H_0 \end{aligned}$$

If we were to use p-values instead of a test statistic, we would find that:

$$\begin{aligned} |t_{stat}| &> t_{\alpha/2,n_1+n_2-2} \iff \text{p-value} < \alpha \\ |t_{stat}| &< t_{\alpha/2,n_1+n_2-2} \iff \text{p-value} > \alpha \end{aligned}$$

Example: Difference in Means - Variances Unknown

Samples of two powders were analysed for their particle diameters. Use a t-test to test whether the powders came from original distributions that had the same or different mean diameters. It is assumed that the values of the population variances σ_1^2 and σ_2^2 are equal. The confidence level should be 95%.

The results of the measurements were

$$\begin{aligned} \text{Sample 1 (11 observations)} &: \overline{x}_1 = 6.65,\ s_1 = 3.91 \\ \text{Sample 2 (16 observations)} &: \overline{x}_2 = 4.28,\ s_2 = 2.83 \end{aligned}$$

The hypothesis are:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

Since is assumed that the values of the population variances σ_1^2 and σ_2^2 are equal we may use the pooled t-test. First we calculate the pooled variance

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(11 - 1)\, 3.91^2 + (16 - 1)\, 2.83^2}{11 + 16 - 2} \\ &= 10.92 \end{aligned}$$

Since the pooled variance is 10.92, the pooled standard deviation is $\sqrt{10.92} = 3.3046$. The test statistic is:

$$\begin{aligned} t_{stat} &= \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{6.65 - 4.28}{3.3046\sqrt{\frac{1}{11} + \frac{1}{16}}} \\ &= 1.831 \end{aligned}$$

We need to check whether $|t_{stat}| > t_{\alpha/2, n_1+n_2-2}$. Since $\alpha = 0.05$, the critical t value is $t_{0.025,25}$, and from the tables, $t_{0.025,25} = 2.06$. Therefore, since $1.831 < 2.06$, we fail to reject $H_0 : \mu_1 = \mu_2$ and state that there is not sufficient evidence at 0.05 significance to conclude that the powders came from original distributions that had different mean diameters. The associated p - value is 0.079 which is greater than the significance level of 0.05.

Effect Size, Sample Size and Statistical Power

In general the effect size represents the magnitude (or strength) of the relationship between two variables. Associated with each effect size is a measure of error (also called reliability or uncertainty), for example, the standard error. In general, **for a sample of a fixed size**, the larger the magnitude of the relationship between variables, the more reliable the relationship i.e. the larger the effect size, the more likely the relationship is to be statistically significant see Figs 3(a) and (b).

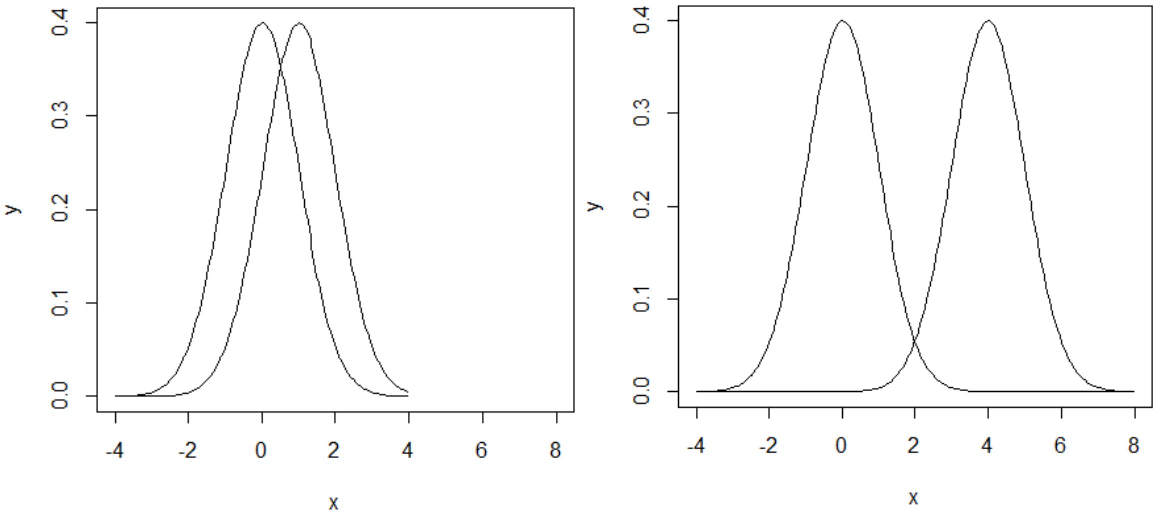


Fig 3(a) A normal distribution with a mean of 0 and a standard deviation of 1 alongside a normal distribution with a mean of 1 and a standard deviation of 1.
Fig 3(b) A normal distribution with a mean of 0 and a standard deviation of 1 alongside a normal distribution with a mean of 4 and a standard deviation of 1.

It is clear from Fig 3. that if we were to sample from each distribution then we are far more likely to find a significant difference between the sample means in Fig 3(b) than Fig.3(a). As you see, the effect size and significance of a relationship are closely related, and we could calculate the significance from the effect size and vice-versa; however, this is true only if the sample size is kept constant, because a relationship of a given strength could be either highly significant or not significant at all, depending on the sample size .

Why does sample size have an effect on the significance of a relationship?

Larger samples increase your chance of finding a significant relationship (if one exists!) because they more reliably reflect the population. The ideal scenario would be one in which we have observations for the whole population in which case any relationships detected would hold. If there are very few observations (N is small), then there are also few possible combinations of the values of the variables, and thus the probability of obtaining **by chance** a combination of those values indicative of a strong relationship is relatively high. Consider the following examples:

Example

Suppose there is an exam that men and women are equally likely to pass. We wish to find out whether gender affects the likelihood of passing the exam so we sample randomly from the population of students who have sat that exam. If there are only four subjects in our sample (two males and two females), then the probability that we will find, purely by chance, a relationship between the two variables can be as high as one-eighth. Specifically, there is a one-in-eight chance that both males will pass the exam and both females will fail, or vice versa (check this).

Example

"Baby boys to baby girls ratio." There are two hospitals: in the first one, 120 babies are born every day, in the other, only 12. On average, the ratio of baby boys to baby girls born every day in each hospital is 50/50. However, one day, in one of those hospitals twice as many baby girls were born as baby boys. In which hospital was it more likely to happen?

.

The Central Limit Theorem

The Central Limit Theorem states that the distribution of the sum (or arithmetic mean) of a large number (say n) of independent, identically distributed variables, with mean μ and variance σ^2 will be approximately normal, regardless of the underlying distribution. Additionally, the Central Limit Theorem tells us that the mean of the distribution (of arithmetic means) is μ and the variance is $\frac{\sigma^2}{n}$.

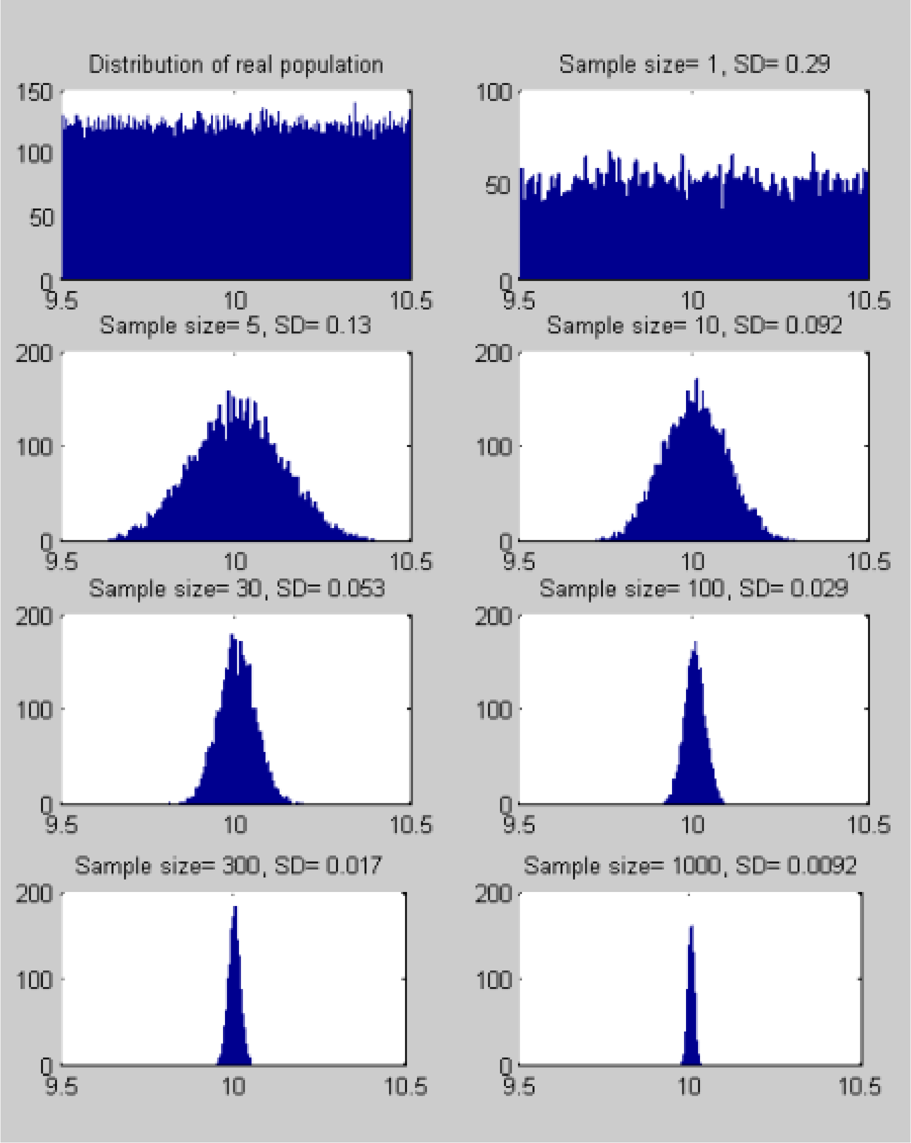


Fig. 4

Expected value of the sample mean

If $X_1, X_2, ..., X_n$ is a random sample of size n drawn from a population (or distribution) with mean μ and variance σ^2.then the expected value (mean) of the sample mean, $E\left[\overline{X}\right] = \mu$.

$$\begin{aligned} E\left[\overline{X}\right] &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n}\left[E\left(X_1\right) + E(X_2) + ... + E\left(X_n\right)\right] \end{aligned}$$

The X_i are identically distributed, which means they have the same mean μ. Therefore, replacing $E\left(X_1\right)$ with the alternative notation μ, we get:

$$\begin{aligned} E\left[\overline{X}\right] &= \frac{1}{n}\left[\mu + \mu + ...\mu\right] \\ &= \frac{1}{n}\left[n\mu\right] \\ &= \mu \end{aligned}$$

We have shown that the expected value (or mean) of the sample mean, $\overline{X}$, is the same as the expected value of the individual X_i.

Variance of the sample mean

Let X_1, X_2, ... , X_n be a random sample of size n from a population (or distribution) with mean μ and variance σ^2, then the variance of the sample mean, $Var\left[\overline{X}\right] = \frac{\sigma^2}{n}$ since:

$$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= Var\left(\frac{X_1}{n} + \frac{X_2}{n} + ... + \frac{X_n}{n}\right) \\ &= Var\left(\frac{X_1}{n}\right) + Var\left(\frac{X_2}{n}\right) + ...Var\left(\frac{X_n}{n}\right) \end{aligned}$$

Recall, $Var\left[X\right] = E\left[(X-\mu)^2\right]$ so $Var\left[\frac{X}{n}\right] = \frac{1}{n^2}Var\left[X\right]$ then

$$Var(\overline{X}) = \frac{1}{n^2}Var\left(X_1\right) + \frac{1}{n^2}Var\left(X_2\right) + ... + \frac{1}{n^2}Var\left(X_n\right)$$

The X_i are identically distributed, which means they have the same variance σ^2. Therefore, replacing $Var(X_i)$ with the alternative notation σ^2 we get:

$$\begin{aligned} Var(\overline{X}) &= \frac{1}{n^2}\left[\sigma^2+\sigma^2+...+\sigma^2\right] \\ &= \frac{1}{n^2}\left[n\sigma^2\right] \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Our result indicates that as the sample size n increases, the variance of the sample mean decreases.

An example of this effect in action is when testing whether there is a difference between two sample means using the two sample t-test . To do this we consider the effect size $(\overline{x}_A - \overline{x}_B)$ and the associated measure of error (or variation), (in this case the pooled standard error). We calculate the associated test statistic:

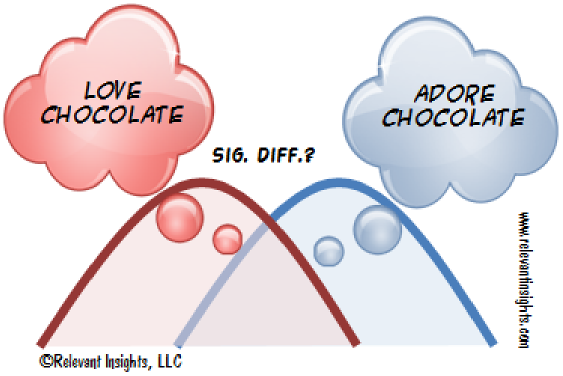
$$t = \frac{\overline{x}_A - \overline{x}_B}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

It is clear from the formula that the larger the sample sizes n_1 and n_2, the smaller the standard error and therefore the larger the test statistic t. The crucial point here is that the larger the sample size, the smaller the error associated with the effect size.

For a data set with very large n, any relationships detected will more than likely be **statistically** significant (i.e. the hypothesis test results in a p-value that is less than the significance level). Bearing this in mind, when inferring relationships from a large data set, it is important to consider the **practical significance** of the result and to do this we look at the effect size. In very large data sets **effect size is far more important than statistical significance** in determining the importance of a relationship.

Example

A large clinical trial is carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistically significant difference in lifespan when using the new treatment compared to the old one. But the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life. Most people would not consider the improvement practically significant.



Measuring Effect Size

In general the effect size represents the magnitude (or strength) of the relationship between two variables. Effect size can refer to a standardised measure of effect (for example Cohen's d statistic) or an unstandardised measure of effect (usually the raw difference). In linear regression, the effect size can be measured by R^2 which is the square of Pearson's correlation coefficient r. The R^2 value measures the proportion of variation in the response variable accounted for by variation in the explanatory variable. Cohen's d statistic measures the standardised difference between two means. It is used in the context of t-tests and is defined to be the difference between two means divided by the pooled standard deviation.

$$d = \frac{\overline{x}_A - \overline{x}_B}{s_p}$$

It has been shown that for small sample sizes, Cohen's d statistic gives a biased estimate of the population effect size. Therefore, Cohen's d is sometimes referred to as the uncorrected effect size. The corrected effect size, or Hedges's g, which is unbiased, is:

$$\text{Hedges } g = d \times \left(1 - \frac{3}{4\,(n_1 + n_2) - 9}\right)$$

It is clear that the effect size is not influenced by the sample size. As well as reporting whether a hypothesis test is statistically significant, it is important to report the effect size and sample size.

Statistical Power

When we perform a hypothesis test, there are four possibilities illustrated in the table below.

	H_0True	H_0 False
Reject H_0	Type 1 Error False Positive α	Correct Inference True Positive $1-\beta$
Fail to reject H_0	Correct Inference True Negative $1-\alpha$	Type II Error False Negative β

The statistical power of a hypothesis test is defined to be the probability of rejecting the null hypothesis H_0 when it is false $(1-\beta)$.

For example, if a study is said to have 80% power then the probability of obtaining a significant result when the effect is real is 0.8. The information shown in the table above is illustrated in Fig. 5 for a two sample t-test.

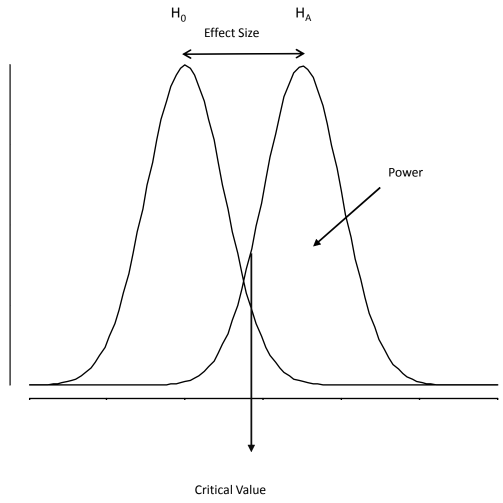


Fig. 5

If the sample size is held constant then a hypothesis test is less likely to detect a significant difference for a small effect size (Fig. 6)

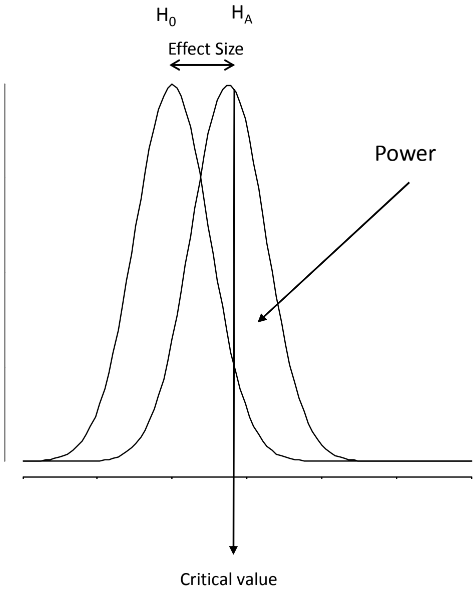


Fig. 6

Before running a study, it is important to run a **power analysis** which allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence. Conversely, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints. In depth knowledge of the area of interest is necessary to determine the minimum effect size that is of practical use.

Bias

> *"Raw Data" is an Oxymoron'* (Lisa Gitelman)

A statistic is **biased** if it is calculated in such a way that is **systematically** (rather than randomly) different from the population parameter of interest. If a data set is not representative of the population then it is said to be biased and any statistics (or estimators) calculated from this data set will be biased. A data set may be biased due to the way that the data is collected.

Example

Measurement bias: an instrument recording temperature is faulty and always records a temperature $0.2°C$ higher than the actual temperature.

Example

Sampling bias: Over 20 million tweets were recorded that referred to hurricane Sandy. The data found that tweets concerning food shopping peaked the night before, and tweets about nightlife peaked just one day after the storm subsided. The data set was biased due to the fact that the majority of the tweets came from Manhattan which was not the most affected area. The parts of New York most affected by the storm were without power and had lower levels of smartphone ownership.

Example

Sampling bias: The Literary Digest voter survey, predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample was not representative of low-income voters, who tended to be Democrats.

Example

The wording in opinion polls can introduce bias e.g. 'Are you in favour of increased taxes?" as opposed to 'Do you wish to see smaller class sizes in schools?'

An estimator, $\widehat{\theta}$ of a population parameter, θ, is said to be unbiased if its expected value, $E\left[\widehat{\theta}\right]$ is equal to the true value of the parameter being estimated i.e. $E\left[\widehat{\theta}\right] = \theta$. For example the sample mean $\overline{X}$ is an unbiased estimator of the true mean μ because $E\left[\overline{X}\right] = \mu$.

The bias of an estimator is the difference between an estimator's expected value and the true value of the parameter being estimated.

$$\text{Bias}\left(\widehat{\theta}\right) = E\left[\widehat{\theta}\right] - \theta$$

Mean Square Error

The value of a predicted estimator, $\widehat{\theta}$, calculated from a sample can differ from the true value of the parameter being estimated, θ, due to **bias** *and* due to the **randomness** in the sample. The mean square error of the predicted estimator $\widehat{\theta}$, measures both the bias and the random variation:

Definition: The Mean Square Error of an estimator (MSE) is

$$MSE(\widehat{\theta}) = E\left[\left(\widehat{\theta} - \theta\right)^2\right]$$

This can be shown to be equal to the square of the expected bias, plus the expected variance:

$$MSE(\widehat{\theta}) = \left[Bias(\widehat{\theta})\right]^2 + Var\left(\widehat{\theta}\right)$$

The probability of a random deviation of a particular size (from the population mean), decreases with an increase in the sample size. Looking at the formula for the MSE of an estimator we see that **an increase in sample size reduces the random variation** associated with the estimator $Var\left(\widehat{\theta}\right)$, but it has **no effect on the bias** associated with the estimator.

The predicted estimator, calculated from samples can differ from the true value of the parameter being estimated due to bias or random variation, Figure 7 below visualizes these differences.

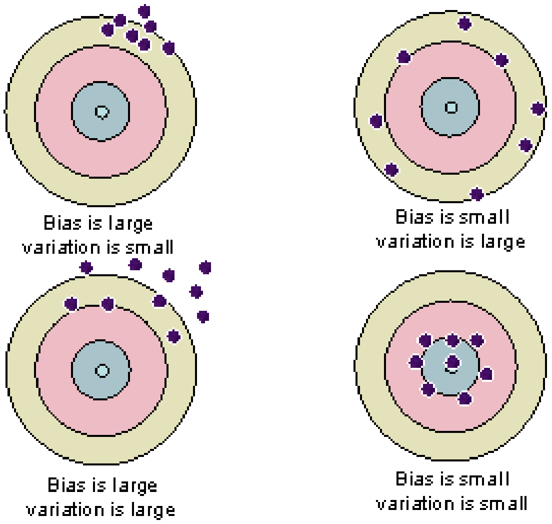


Fig. 7 The effects of bias and random variation on predictions.

In large data sets we have small variation but may be susceptible to bias, represented by the target in the top left of Figure 8. It is important to mindful of this when analysing large data sets, an abundance of statistically significant results may encourage us to be overconfident in the predictive ability of our analysis.

> *'Even if the amount of knowledge in the world is increasing, the gap between what we know and what we think we know may be widening'* 'Nate Silver, The Signal and the Noise