Frequentist Inference, Interval Estimation and Hypothesis Testing

CMED6040 – Session 1

Tim Tsang (matklab@hku.hk)

School of Public Health
The University of Hong Kong

9 May 2023

Course format and assessment

- Format: lecture + practical
  - approximate time allocation: 90 + 60 min
- Tutorial after several lectures
- Coursework: 50% - 3 assignments
- Final exam: 50%
  - August 1, 2023 (Tuesday)
  - 18:30-20:30
  - Open book
- Grading: high low

High:
- Appropriate analytic method
- Accurate numerical results
- Clear presentation of the results and choice of methods
- Interpretation of the results relevant to the public health context

Low:
- Unclear / wrong use of analytic method
- Inaccurate numerical results
- Poor presentation
- No interpretation of the results

Session 1 outline

- 18:30 to 19:00 – Frequentist inference
- 19:00 to 19:30 – Interval estimation
- 19:30 to 20:00 – Hypothesis testing
- 20:00 to 21:00 – Practical

Session 1 learning objectives

After this session, students should be able to

- Define the likelihood function and calculate the likelihood for simple probability models
- Interpret parameter estimates and confidence intervals
- Calculate and interpret p-values for simple hypothesis tests

From probability to inferential statistics

A simple example

Consider the following example. The parameter φ is unknown but can have two possible values, 0 or 1. The variable X can be observed, and depends on φ.

- If $\varphi = 0$ then $X = 0$ with probability 5 / 6, and $X = 1$ otherwise.

- If $\varphi = 1$ then $X = 0$ with probability 1 / 5, and $X = 1$ otherwise.

How does observation of X help us to estimate φ?

A simple example

The four possibilities

X	φ	
	0	1
0	5 / 6	1 / 5
1	1 / 6	4 / 5

Likelihood – formal definition

The likelihood function of a parameter θ is the function that associates $p(x|\theta)$ to each θ. Formally,

$$l(\theta \mid x) = p(x \mid \theta)$$

- Larger values of l indicate that the event under consideration is more likely for that particular value of θ.
- For fixed (observed) x we use the likelihood function to determine the plausibility (or *likelihood*) of each value of θ.
- We can estimate θ as the value of θ that maximizes $l(\theta \mid x)$.

A second example

We toss a coin of indeterminate fairness twice. The (unknown) probability that it will come up heads (H) is written θ. The number of heads is written *X*, so *X* can take values 0, 1 or 2. *X* follows a Binomial(2, θ) distribution,

- $\Pr(X = 0 \mid \theta) = (1 - \theta)^2$
- $\Pr(X = 1 \mid \theta) = 2\theta(1 - \theta)$
- $\Pr(X = 2 \mid \theta) = \theta^2$

How does observation of *X* help us to estimate θ?

Likelihood function for x = 1

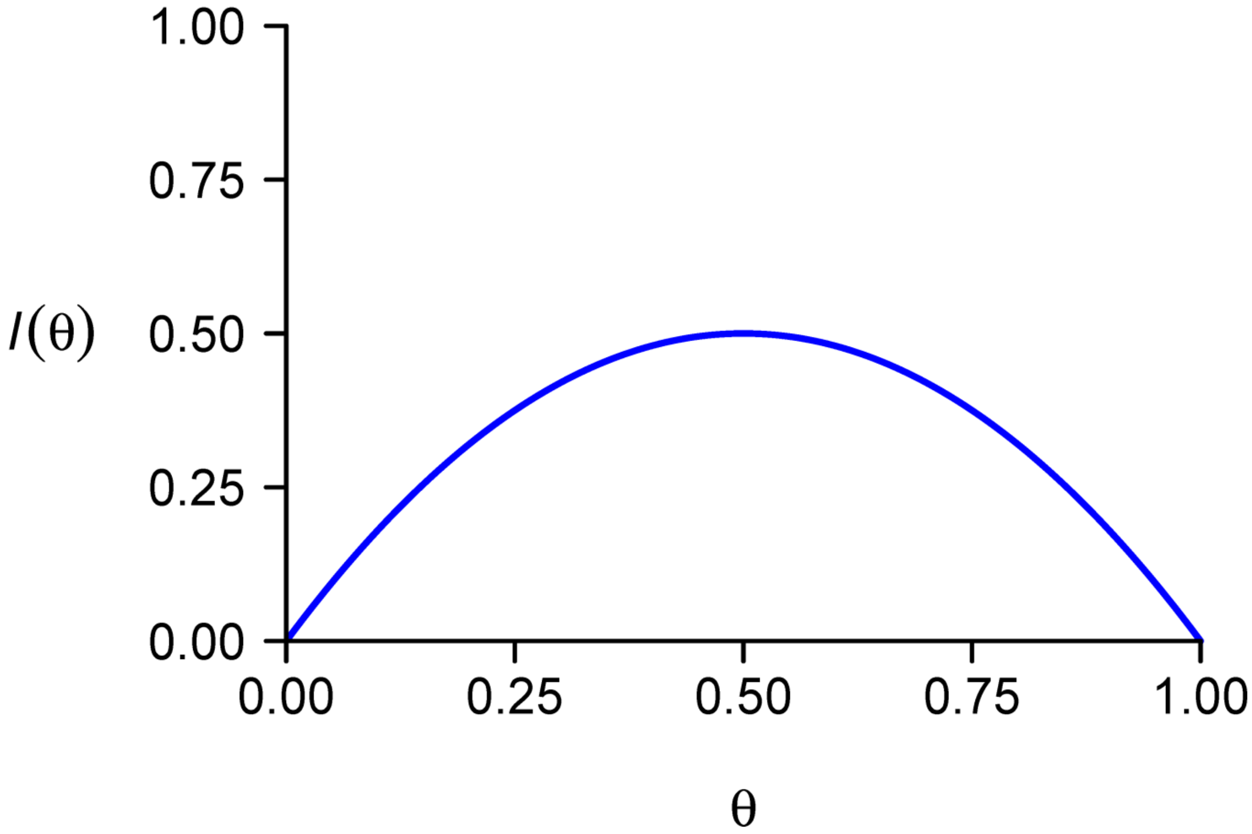


Figure: If x = 1, $l(\theta \mid x = 1) = 2\theta(1 - \theta)$ and the value of θ with highest likelihood is 1 / 2

Likelihood functions for different values of x

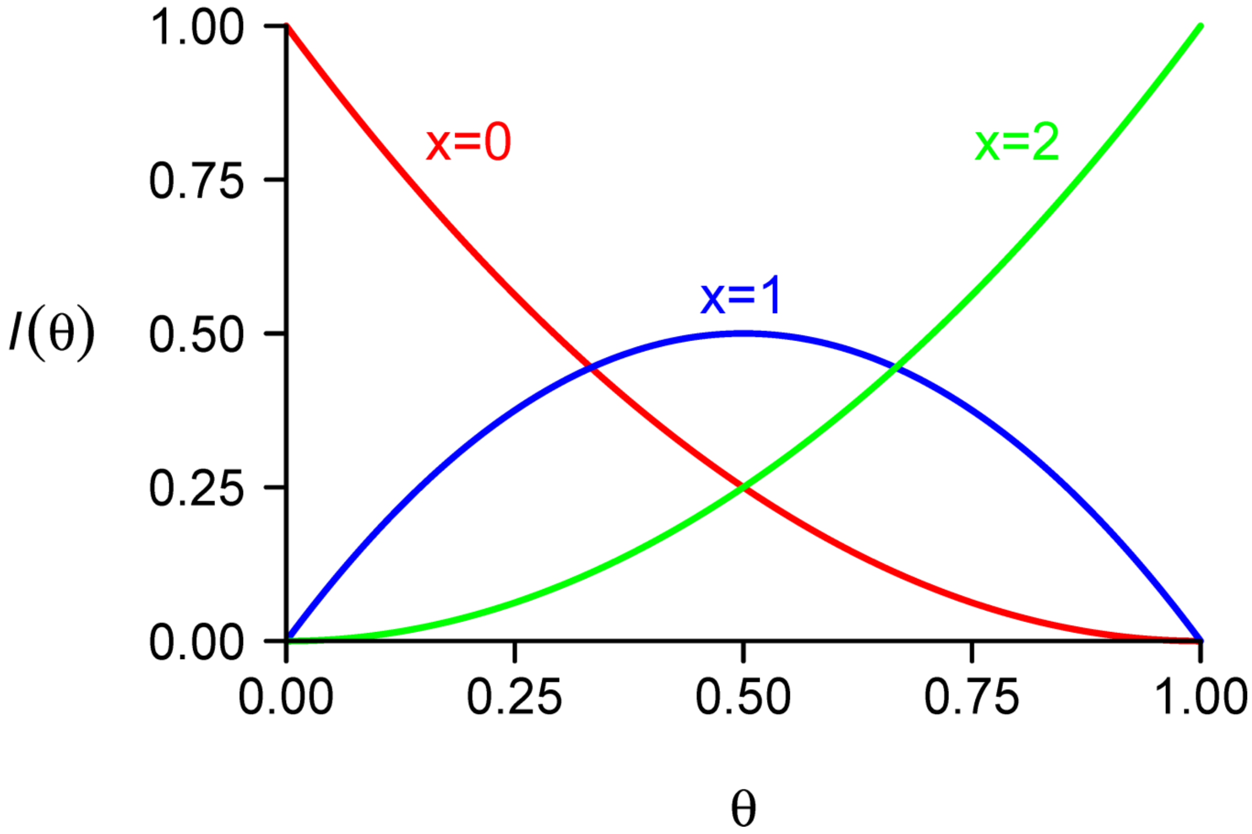


Figure: Likelihood functions for x = 0, 1, 2 where the most likely values of θ are 0, 0.5 and 1 respectively.

A third example

We observe a random variable Y which follows a Poisson distribution with rate λ:

- $\Pr(Y = 0 \mid \lambda) = \exp(-\lambda)$
- $\Pr(Y = 1 \mid \lambda) = \lambda \exp(-\lambda)$
- $\Pr(Y = 2 \mid \lambda) = \lambda^2 \exp(-\lambda)/2$
- $\cdots$
- $\Pr(Y = y \mid \lambda) = \lambda^y \exp(-\lambda)/y!$

How does observation of Y help us to estimate λ?

Likelihood functions for different values of *y*

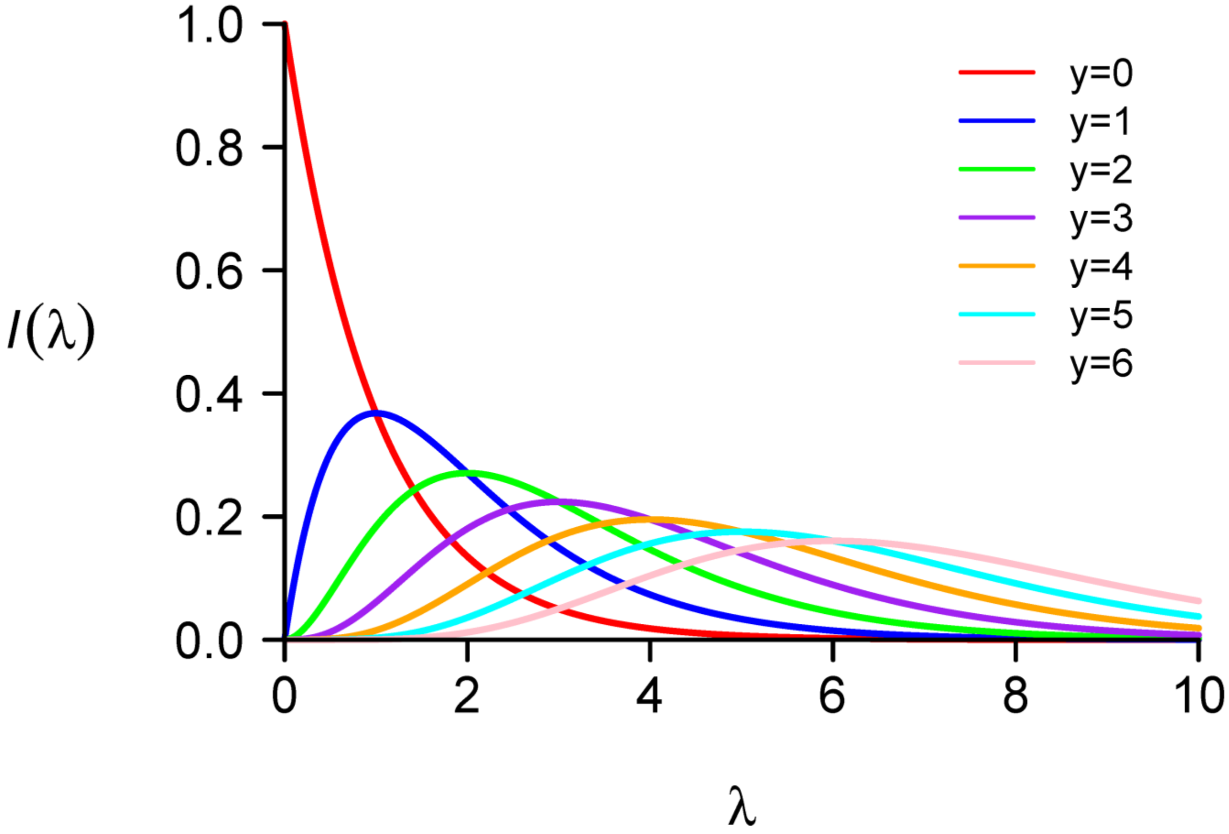


Figure: Likelihood functions for $y = 0, 1, 2, \ldots, 6$.

Likelihood – proportional definition

In many problems it is sufficient to define the likelihood function in a more general sense as

$$l(\theta \mid x) \propto p(x \mid \theta)$$

- The 'constant of proportionality' isn't important if we are interested in maximizing the likelihood
  - For example, whichever value of θ maximizes $\theta(1-\theta)$ will also maximize $k\theta(1-\theta)$
- In other cases, we may be able to get back the constant of proportionality later if we need it.

Multiple observations

- Recall that if A and B are independent,

$$\Pr(A \cap B) = \Pr(A, B) = \Pr(A)\Pr(B)$$

- In the same way, if observations x_1 and x_2 can be considered independent, then $p(x_1, x_2 \mid \theta) = p(x_1 \mid \theta)\ p(x_2 \mid \theta)$.
- So $l(\theta \mid x_1, x_2) = l(\theta \mid x_1)\ l(\theta \mid x_2)$.

- In general for observations $\boldsymbol{x} = x_1, \ldots, x_n$

$$l(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} l(\theta \mid x_i) = l(\theta \mid x_1)\ l(\theta \mid x_2) \ldots l(\theta \mid x_n)$$

Log-likelihood function

- Sometimes the likelihood can be very small, because it is a product of very small numbers
- This can make its computation unstable

- The logarithm of the likelihood function has the same maximum as the likelihood function, and has nicer computational properties, so we can use that instead to help estimate parameters
- The logarithm of the likelihood is usually called the “log-likelihood” function

Maximum Likelihood Estimator (MLE)

- Maximum likelihood estimator is obtained by maximizing the (log)-likelihood function

- MLE has some good properties
- Consistency: when the sample size increases, MLE will approach the true unknown value (at any precision)
- Asymptotic normality: with a large sample size, MLE is normally distributed

Example – asthma prevalence

In a recent survey, 60 of 568 sampled children in Hong Kong were found to have asthma. Data are available on the number of cases x_i in n_i sampled children in each of i = 1, …, 18 districts.

- Suppose probability θ describes risk of asthma.

- Likelihood contribution for each district is proportional to

$$\theta^{x_i}\,(1-\theta)^{n_i-x_i}$$

- Likelihood function is proportional to

$$\prod_{i=1}^{18} \theta^{x_i}\,(1-\theta)^{n_i-x_i} = \theta^{\sum_{i=1}^{18} x_i}\,(1-\theta)^{\sum_{i=1}^{18} n_i-x_i}$$

- Then $l_{all}(\theta \mid \boldsymbol{x}) \propto \theta^{60}\,(1-\theta)^{508}$

The log-likelihood function

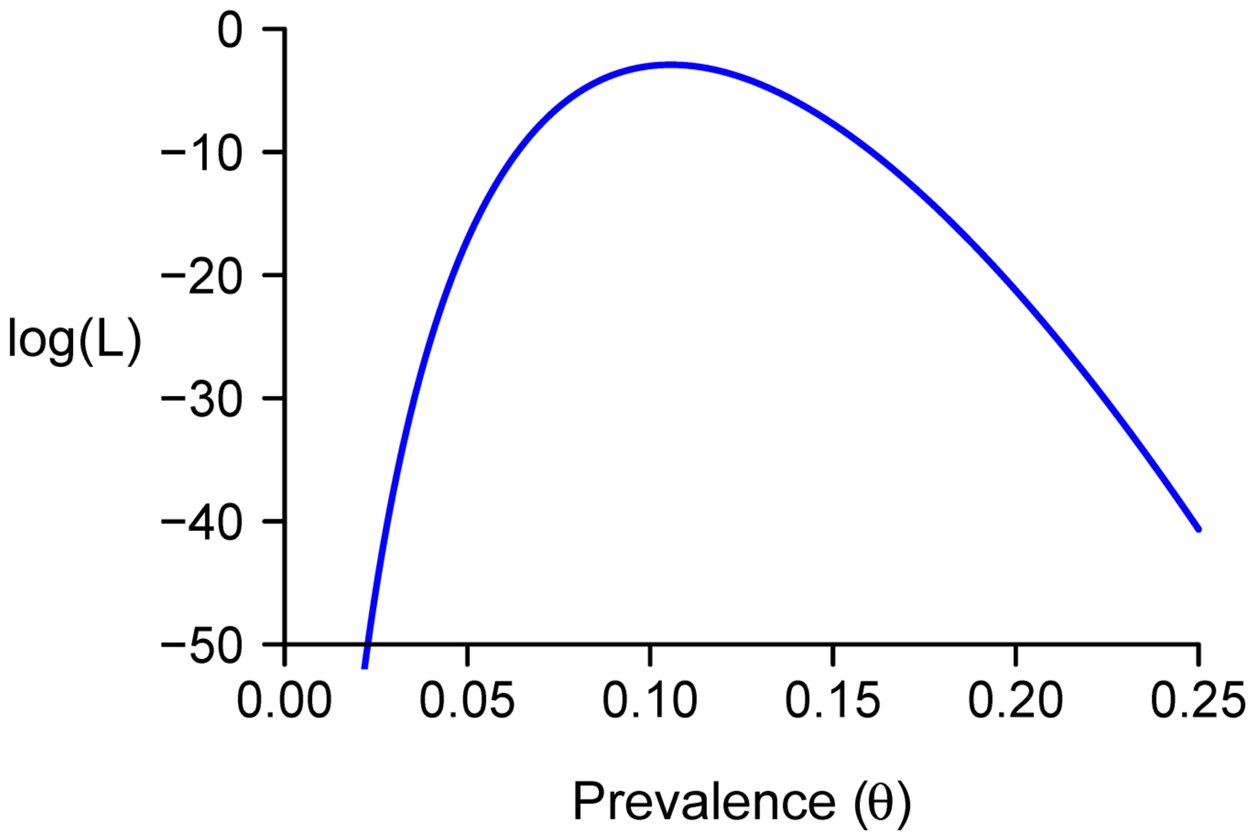


Figure: Log-likelihood function for asthma prevalence.

Example – influenza incidence

Number of children x_i admitted to public hospitals in March 2007 and subsequently diagnosed with influenza, from each of 18 districts with n_i children in each district.

- The simple Poisson model assumes that each x_i follows a Poisson distribution with constant mean λn_i, so

$$p(x_i \mid \lambda) = \frac{(\lambda n_i)^{x_i} \exp(-\lambda n_i)}{x_i!}$$

- Then the likelihood for any particular mean λ given an observed value of x_i is $l_b(\lambda \mid x_i) \propto \lambda^{x_i} \exp(-\lambda n_i)$

The Poisson likelihood

- With 330 cases of influenza in a total population of 720,100 children, the likelihood function in this example is

$$
\begin{aligned}
l_b(\lambda \mid x_i) &\propto \prod_{i=1}^{18} \lambda^{x_i} \exp(-\lambda n_i) \\
&\propto \lambda^{\sum_{i=1}^{18} x_i} \exp\left(-\lambda \sum_{i=1}^{18} n_i\right) \\
&\propto \lambda^{330} \exp(-720100\lambda)
\end{aligned}
$$

where $\boldsymbol{x} = \{x_1, \ldots, x_{18}\}$.

The log-likelihood function

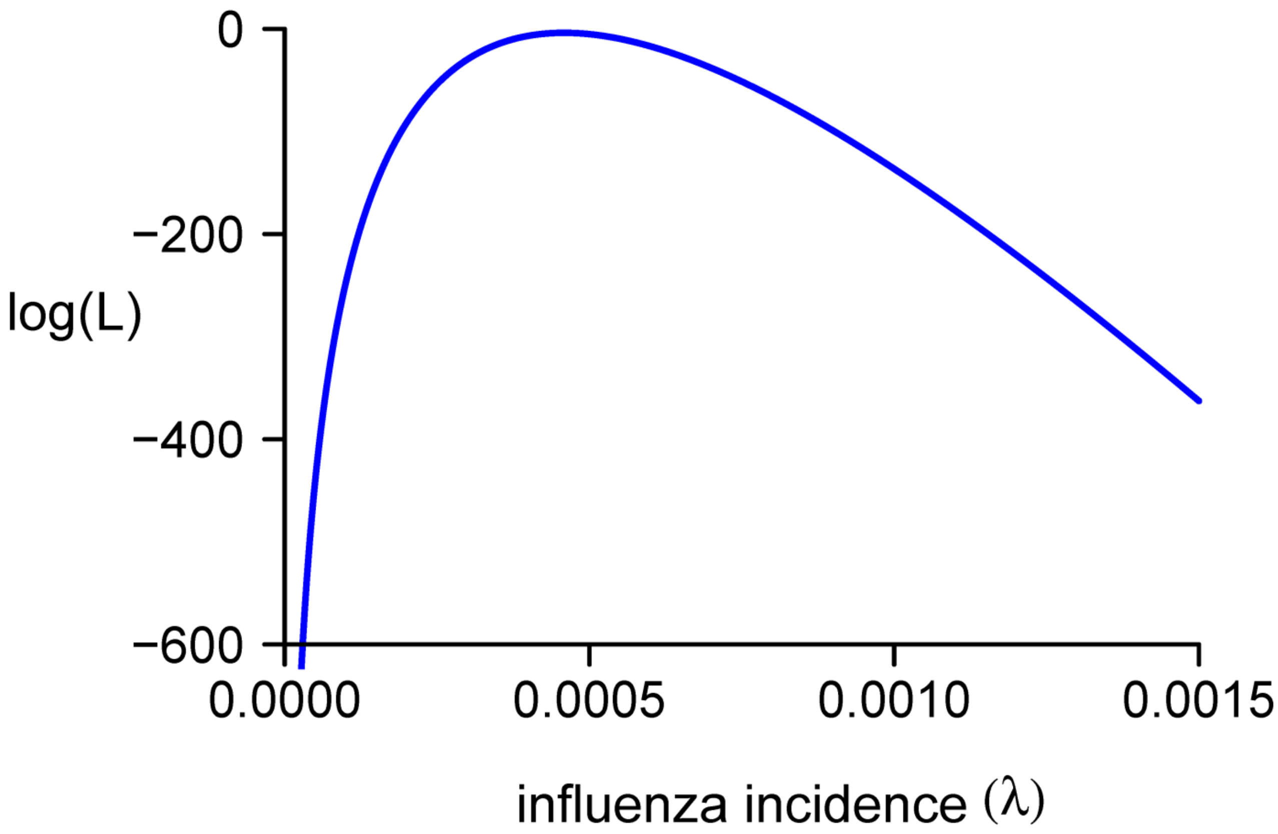


Figure: Log-likelihood function for influenza incidence.

Allowing for vaccination

To estimate the effect of vaccination, we can allow the parameter of the Poisson distribution to depend on a vaccination policy indicator z_i (1=yes, 0=no).

- We can specify $\lambda_i = \exp(\beta_0 + \beta_1 z_i)$
- Then the likelihood for $\boldsymbol{\beta} = \{\beta_0, \beta_1\}$ given an observed value of x_i and z_i is

$$l_c(\beta |\, x_i, z_i) \propto {\lambda_i}^{x_i} \exp(-\lambda_i n_i)$$

$$\propto \exp(\beta_0 + \beta_1 z_i)^{x_i} \exp(-\exp(\beta_0 + \beta_1 z_i)\, n_i)$$

The likelihood function allowing for vaccination

- With 147 vaccinated cases and 434,900 children in vaccination districts, the likelihood function for this example is now

$$l_c(\beta \mid \boldsymbol{x}, \boldsymbol{z}) \propto \prod_{i=1}^{18} \lambda_i^{x_i} \exp(-\lambda_i n_i)$$

$$\propto \prod_{i=1}^{18} \exp(\beta_0 + \beta_1 z_i)^{x_i} \exp(-\exp(\beta_0 + \beta_1 z_i)\, n_i)$$

$$\propto \exp(330\beta_0 + 147\beta_1) \times \exp(-285200 \exp(\beta_0) - 434900 \exp(\beta_0 + \beta_1))$$

Interval estimation

The normal distribution

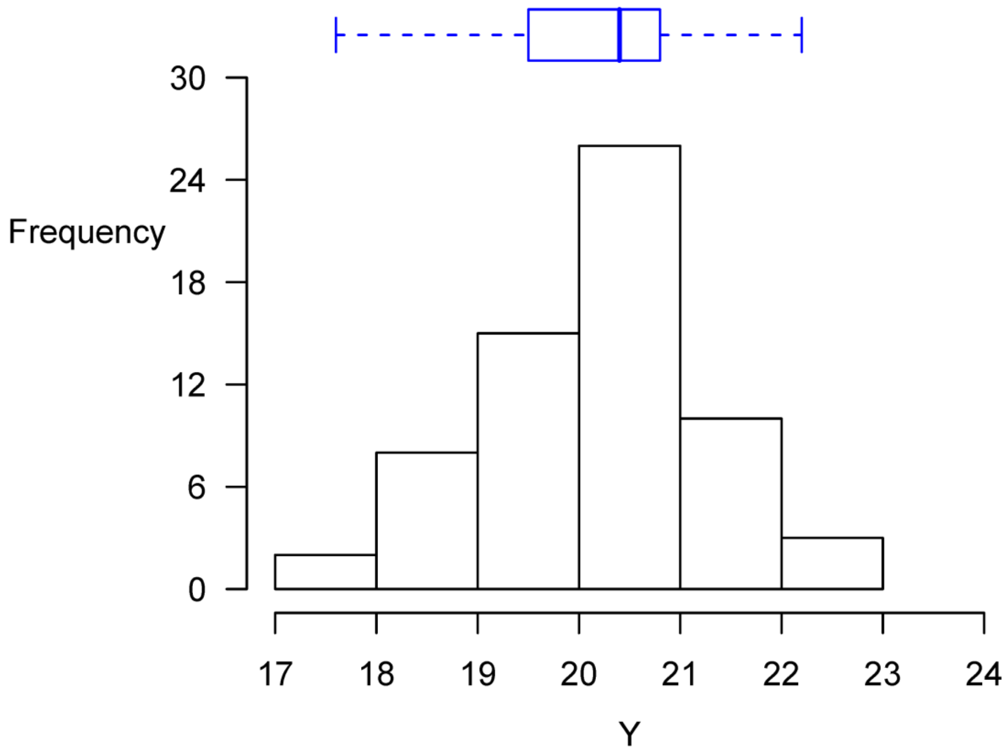
Sample from a normal distribution with

```
rnorm(n, mean = 0, sd = 1)
```

- *n* is the number of samples
- *mean* is the mean
- *sd* is the standard deviation

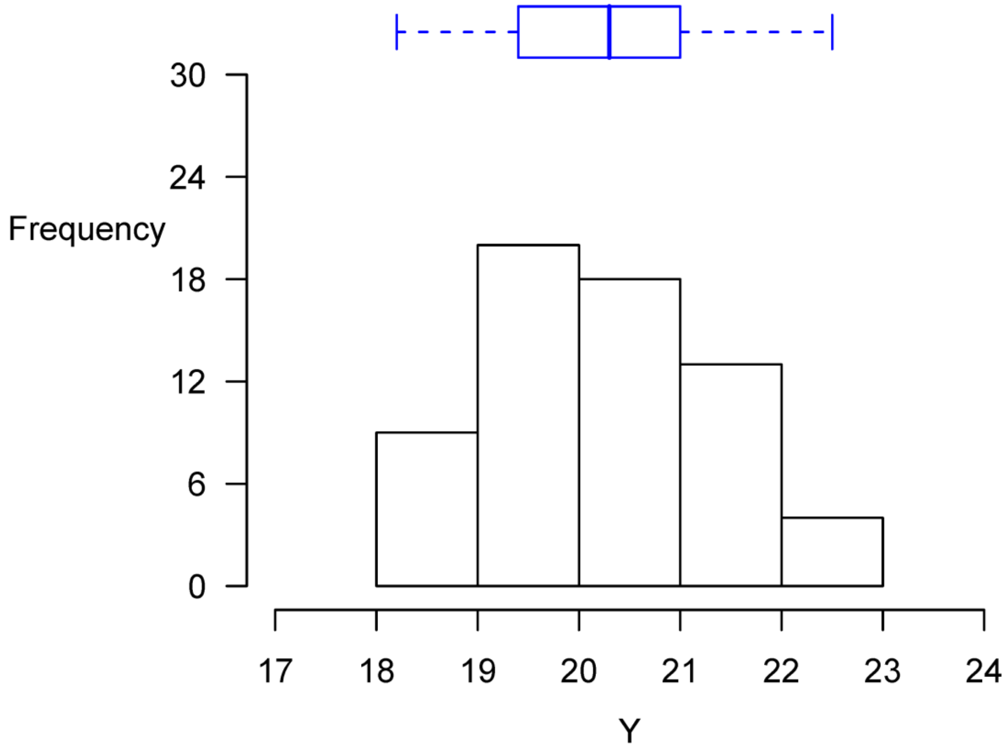
The normal distribution

- If we sample 64 times from a N(20, 1) distribution using `rnorm(64, 20, 1)`, the sample might look like this:



The normal distribution

- Or it might look like this:



Normal distribution

- Different data can arise from identical probability distributions, and identical data can arise from different probability distributions.

- Finding out which probability distribution led to the data is therefore not trivial.

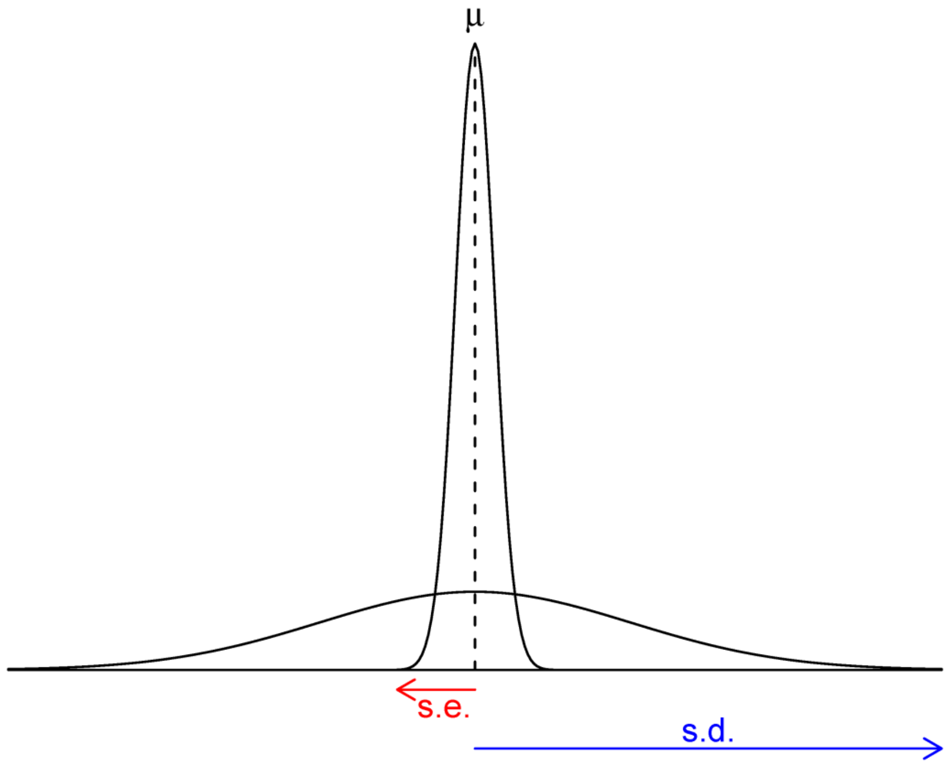
Means of repeated samples

- If we have a single sample of size 64 from a Normal (μ,1) distribution, the best estimate of the mean μ is the sample mean $\bar{x}$.

- According to the Central Limit Theorem, under repeated sampling the sample means will follow a normal distribution with mean μ and standard error of the mean $\sigma/\sqrt{n}$
- i.e., 1 / 8 in our example since σ = 1 and n = 64.

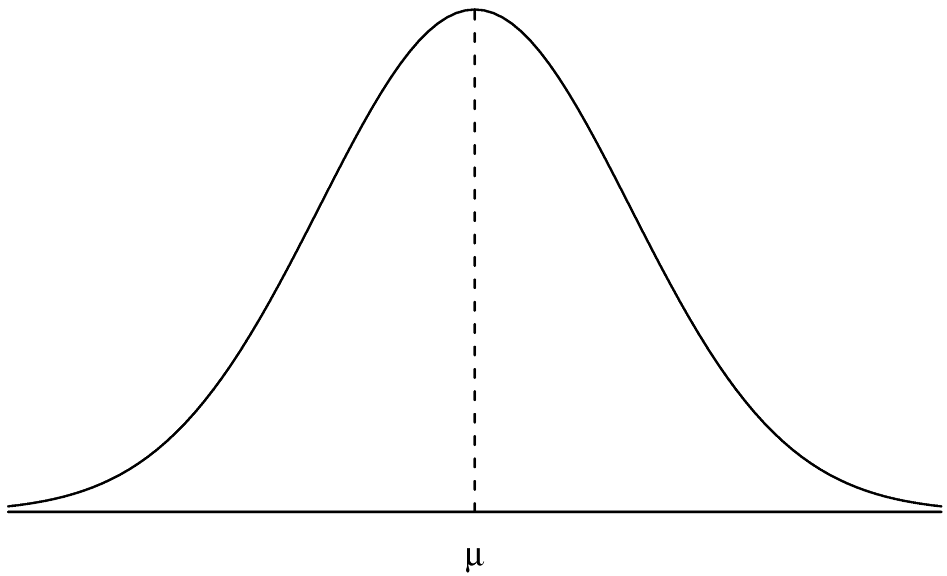
Standard error versus standard deviation

Figure: Distribution of sample means vs distribution of sample. The standard deviation (σ) refers to the spread of data. The standard error $\sigma/\sqrt{n}$ refers to the variability of the mean under repeated sampling.



CI for the sample mean

- Recall the Central Limit Theorem: If *X* follows a distribution with mean μ and standard deviation σ, and we take a random sample of size *n*, provided that *n* is sufficiently large the sample mean $\bar{X}$ will follow a normal distribution, $\bar{X} \sim Normal(\mu,\ \sigma^2/n)$.
- Hence if we drew repeated random samples of size n from a population, 95% of the $\bar{X}$s we will see will fall within $\mu \pm 1.96\sigma/\sqrt{n}$.



Possible case for $\bar{X}$

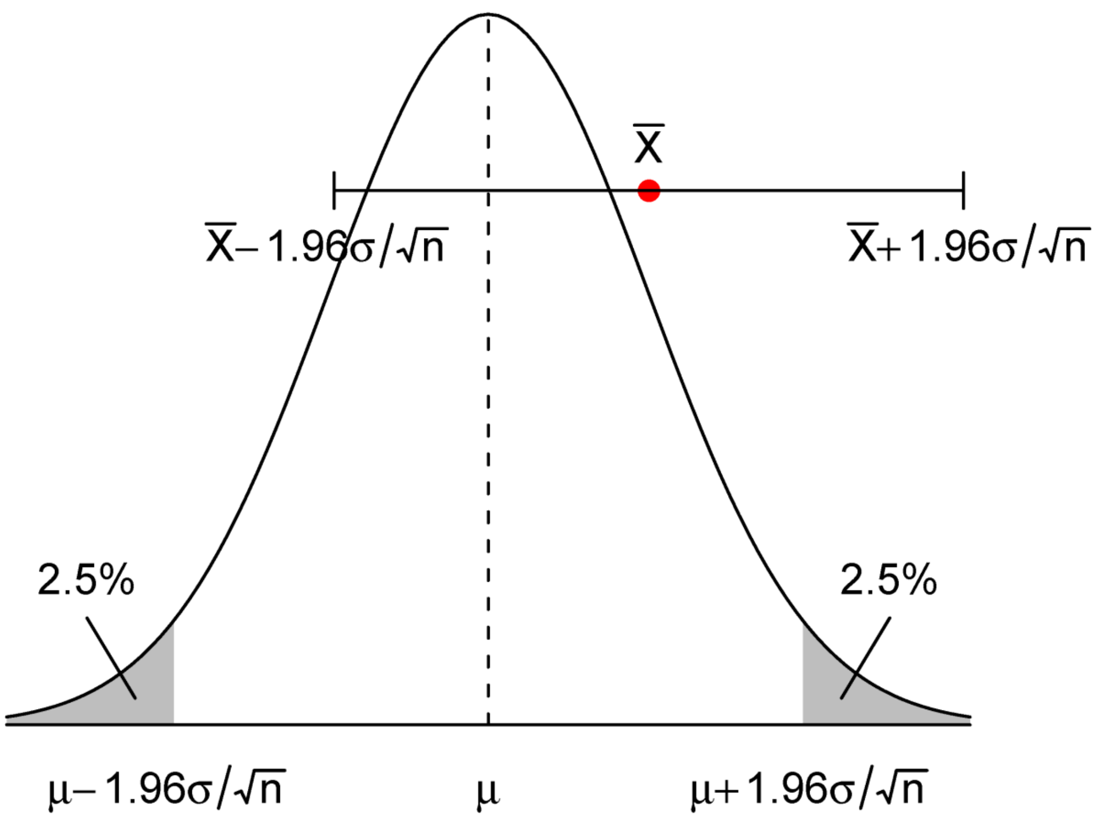


Figure: $\bar{X}$ could be sampled here.

Possible case for $\bar{X}$

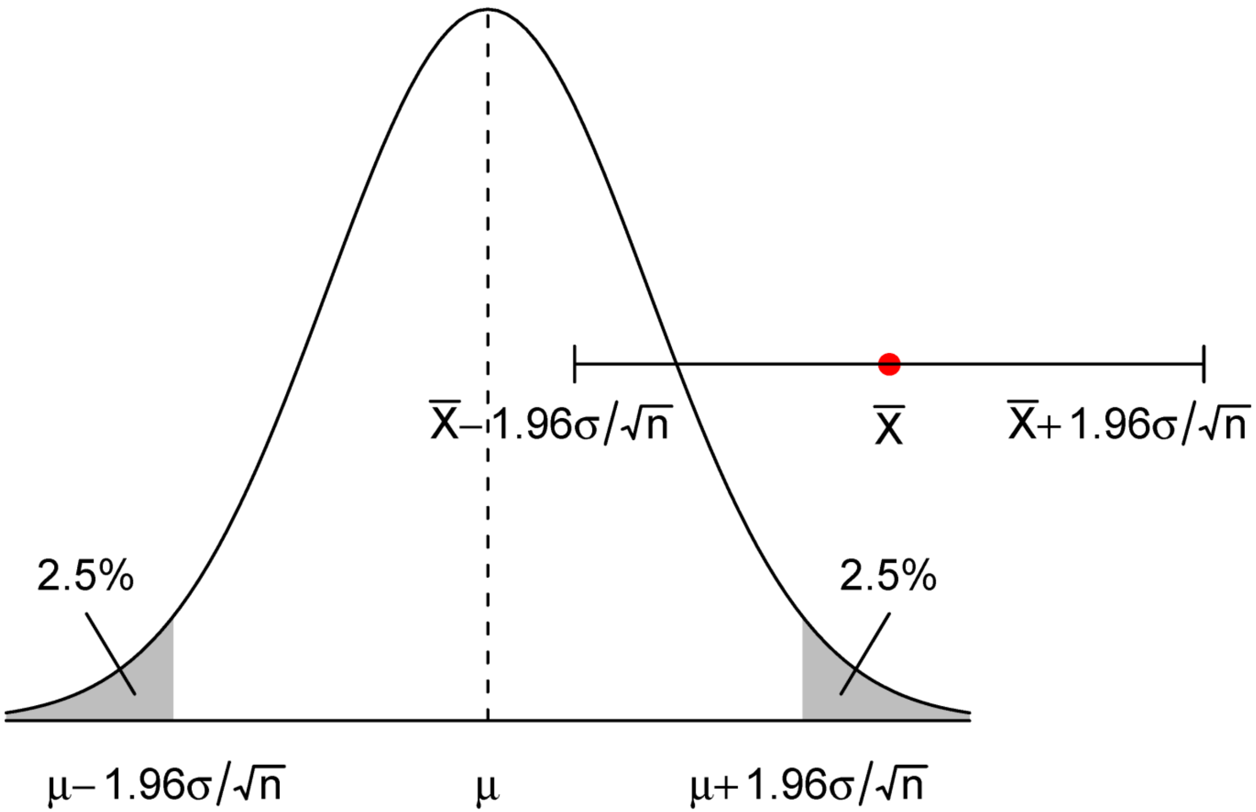


Figure: In 5% of samples $\bar{X}$ will be in the tails of the distribution and then the 95% CI will not include μ

Definition of a confidence interval

- Under repeated samples, we can say that P% of P% confidence intervals will contain the true population value.
  - For example, 95% of all 95% confidence intervals will cover the true population value.

- A single CI may or may not cover the true value.

- We can say that we have 95% confidence that a single 95% CI will cover the true value, but this is simply a short version of the definition above.

- Strictly speaking, we cannot say that there is a 95% probability that a single 95% CI will cover the true value.

Derivation of confidence intervals

- If the sampling distribution of the parameter of interest is known, it is straightforward to calculate a confidence interval.

- e.g. if the sampling distribution of a parameter θ follow a Normal distribution, we can use $(\hat{\theta} - z_{\frac{\alpha}{2}}\sigma/\sqrt{n}, \hat{\theta} + z_{\frac{\alpha}{2}}\sigma/\sqrt{n})$.

- 95% confidence intervals for regression coefficients in linear regression models are typically estimated this way.

Example – influenza admissions (Poisson)

- Number of admissions X follows a Poisson distribution with mean μ. We set $\mu = \lambda \times n$ where n was the number of children (population) and λ was the admission rate.
- Using the Normal approximation and remembering that the variance of a Poisson is the same as the mean,

$$X \sim Normal(\mu, \mu)$$

- A 95% confidence interval for μ is therefore

$$\bar{X} - 1.96\sqrt{\bar{X}}, \bar{X} + 1.96\sqrt{\bar{X}}$$

Example – influenza admissions (Poisson)

- In the data, $\bar{X} = 330, n = 720100$, so a 95% CI is

  $$\left(330 - \ 1.96\sqrt{330}, 330 + 1.96\sqrt{330}\right)$$

  which is (294.4, 365.6)

- We are more interested in a 95% CI for λ and since *n* is fixed. We can simply use the transformation $\lambda = \mu/n$ to find the new CI.
- A 95% CI for λ is therefore

  $$\left(330 - \ 1.96\sqrt{330}, 330 + 1.96\sqrt{330}\right)/720100$$

  which is (0.000409, 0.000507). The MLE is $\widehat{\lambda}$ = 0.000458.

A simple method of deriving confidence intervals

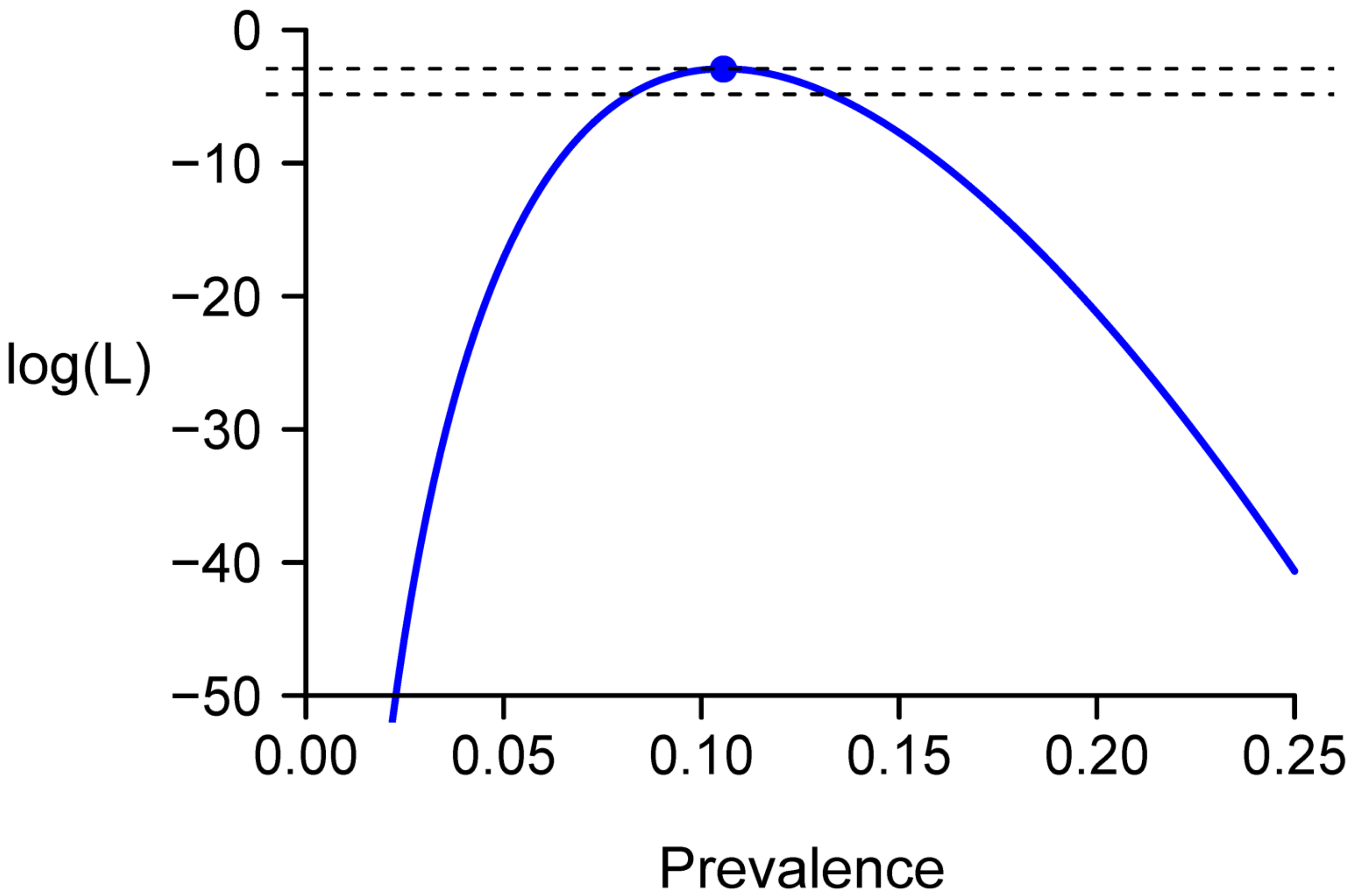
- Another method to derive confidence intervals is based on likelihood ratios (Wilks' theorem), using the (asymptotic) property that

$$2\left|\log l(\hat{\theta}) - \log l(\theta)\right| \sim \chi_1^2$$

- Note that the 95th percentile of the χ_1^2 distribution is 3.84.
- Therefore the values of θ for which $\left|\log l(\hat{\theta}) - \log l(\theta)\right| < 1.92$ will form a 95% confidence interval for θ.

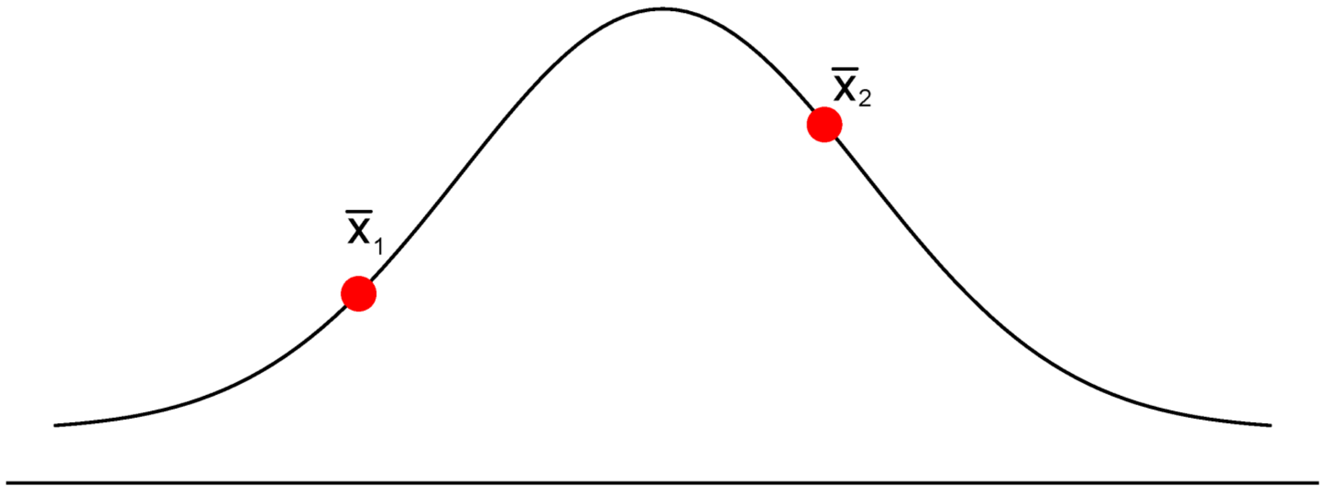
- Warning – may not work very well if parameters are correlated with each other.

Asthma example



Hypothesis testing

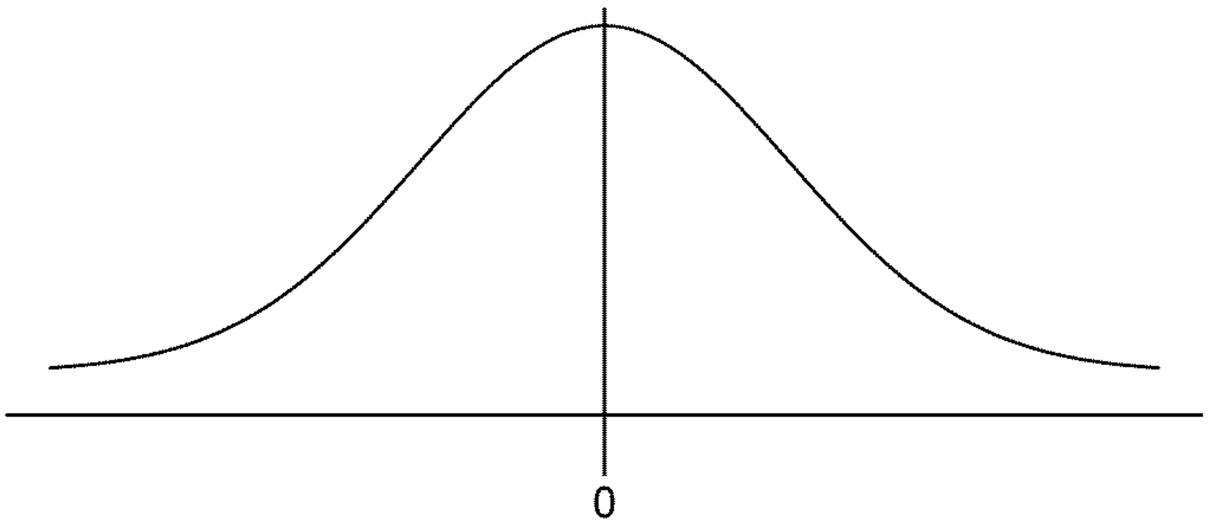
Comparing groups



Null hypothesis – assume both groups are sample from distributions with the same mean. What is the chance of getting a difference $\bar{x}_1 - \bar{x}_2$as usual or more unusual than the difference observed?

Comparing groups

Under the null hypothesis, $\bar{x}_1 - \bar{x}_2$ will have a Normal distribution with mean 0 and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.



UN Survey

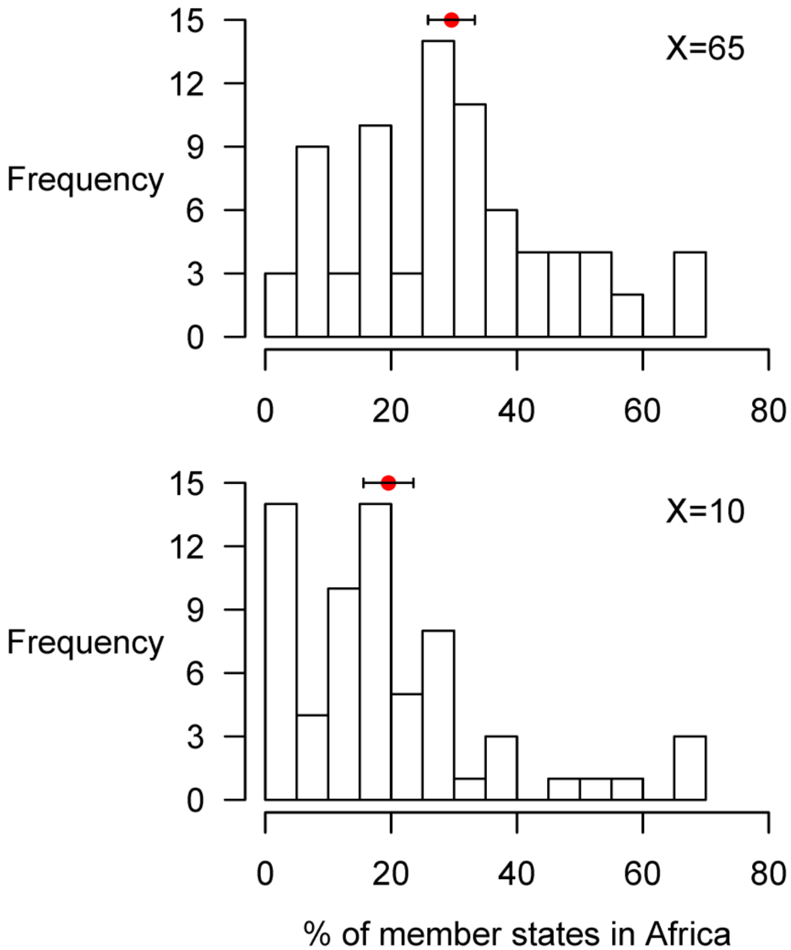


Figure: Responses of 77 students given $X = 65$, and 65 students given $X = 10$.

Observed difference versus sampling distribution

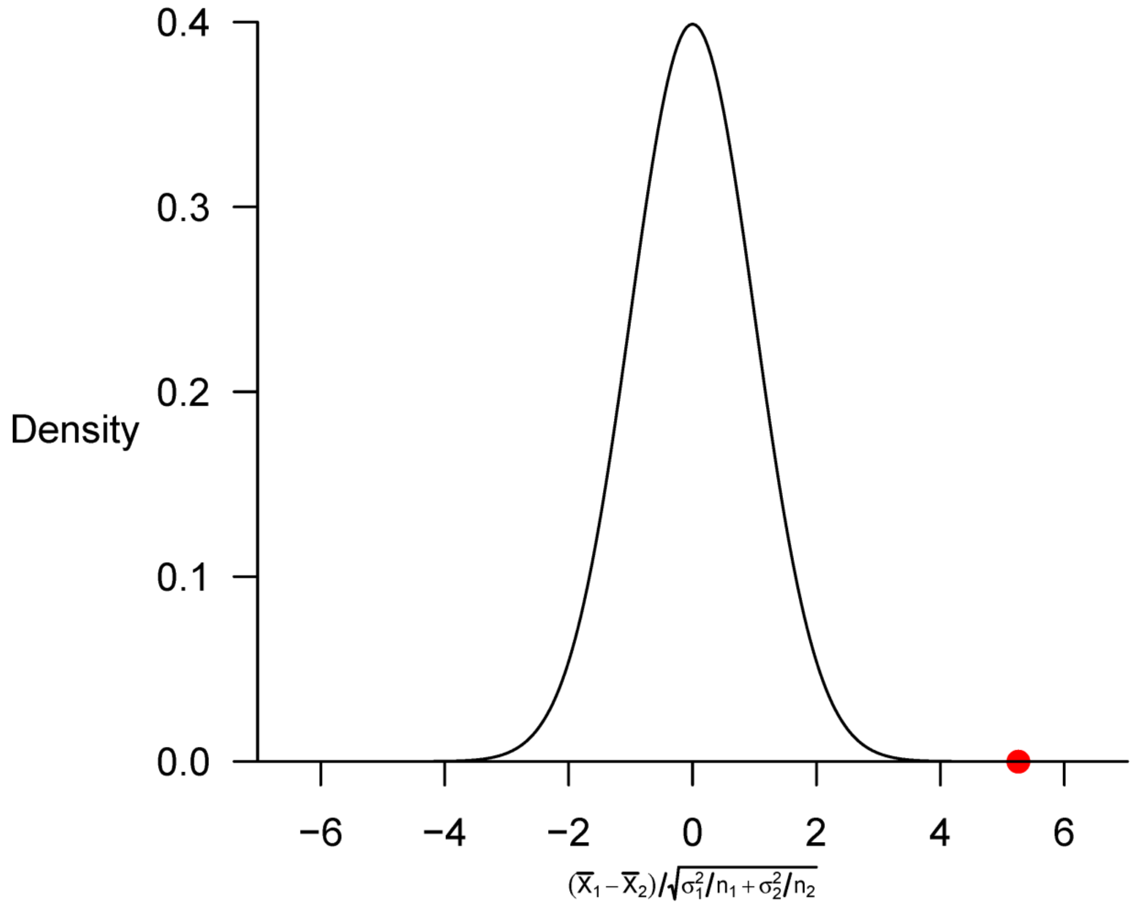


Figure: An observed standardized difference of 5.26 is at the extremes of the sampling distribution under the null hypothesis.

Plausibility of results under the null hypothesis

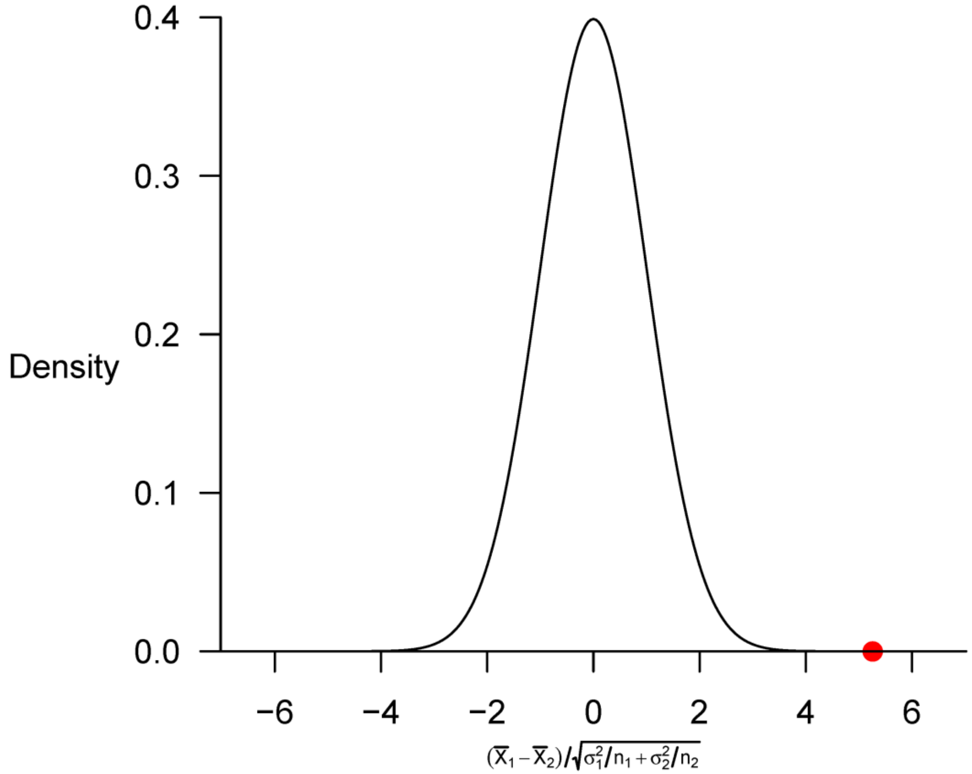


Figure: If the null hypothesis were true, i.e. no difference between means, it would be very unusual to observe such a large difference (whether less than –5.26 or greater than 5.26). We would only observe such a large difference in 1% of repeated experiments.

How do we interpret this?

- If we repeated this experiment many times, and *if* the null hypothesis were true, we would only see differences greater than 5.26 (or less than –5.26) in 1% of those experiments.
- The value of 0.01 or 1% is often referred to as a p-value
- Notice that the p-value is a conditional probability – it is conditional on the null hypothesis being true.
- Small p-values, indicating that observed differences are unlikely under the null hypothesis, are usually taken as evidence *against* the null hypothesis
- A common threshold is $p < 0.05$; in that case p-values less than 0.05 are called ‘statistically significant’.

Common misunderstandings about p-values

1. The p-value is not the probability that the null hypothesis is true.

- The p-value is *p*(*such unusual data* | *null hypothesis is true*), whereas the probability that the null hypothesis is true is *p*(*null hypothesis is true* | *such unusual data*).
- We cannot derive the second probability without some assumption about *p*(*null hypothesis is true*)

Common misunderstandings about p-values

2. The p-value is not the probability that a finding is “merely due to chance”.

- As the calculation of a p-value is conditional on the assumption that a finding is the product of chance alone, it cannot simultaneously be used to gauge the probability of that assumption being true.
- The p-value is the probability that a finding is “merely due to chance” *if* the null hypothesis is true.

Common misunderstandings about p-values

3. The p-value does not indicate the size or importance of the observed effect (compare with effect size).

- In a large sample, the standard errors will be small, and therefore even small differences may be associated with small p-values.

Common misunderstandings about p-values

4. A p-value of 1.00 does not mean the null hypothesis is true

- A p-value of 1.00 indicates that the observed data were completely consistent with no effect (for example the primary outcome occurred at exactly the same rate in two groups)
- Study of 10 people, Group A: 2/5 vs Group B: 2/5 experience the event of interest – p-value for difference = 1.00
- Study of 1000 people, Group A: 200/500 vs Group B: 200/500 experience the event of interest – p-value for difference = 1.00