THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Chapter 7–9

- Statistical Inference
- Central Limit Theorem
- Confidence Intervals
- Hypothesis Testing

Statistical inference

Statistical Inference is the process of forming judgements about the parameters.

Assumptions:

- Assume that data $X_1, \ldots, X_n$ is drawn randomly from some ***unknown*** distribution (identically distributed).

- Assume that the data is independent

  $\longrightarrow X_i$ are i.i.d. (independent and identically distributed), i.e.,

  1. $X_i \sim G$ for all $1 \leq i \leq n$
  2. X_is are independent

A statistic

A **statistic** is any function of the observations in a random sample.

$\longrightarrow$ A statistic is itself a R.V.

Examples:

- $g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ = Sample mean
- $g(X_1, X_2, \ldots, X_n) = \max\{X_1, X_2, \ldots, X_n\}$
- Sample variance and sample standard deviation
- Sample quantiles besides the median, (quartiles and percentiles)

Some notations:

- The probability distribution of a statistic is called the **sampling distribution**.
- A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.
- The statistic $\hat{\Theta}$ is called the point estimator.

Example:

Sample Mean $= \overline{X} =$ estimator of the population mean, μ.

Normal Distribution - Recap

$X \sim N(\mu, \sigma^2)$ then pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

- $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$
- If $\mu = 0$ and $\sigma = 1$ then

  $$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R},$$

  = standard normal distribution
- $\frac{X-\mu}{\sigma} \sim \mathsf{N}(0,1)$ = standardization
- $X = \mu + \sigma Z, \quad Z \sim \mathsf{N}(0,1)$

Central Limit Theorem (for sample means)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and finite variance σ^2,then

$$\lim_{n\to\infty} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1)$$

where $\bar{X}$ is the sample mean. Equivalently,

$$P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x)$$

Regardless of X_i's distribution, the sum behaves (approximately) as the Gaussian random variable!

$$\bar{X} \quad \overset{n\to\infty}{\approx} \quad N\left(\mu, \frac{\sigma^2}{n}\right)$$

$S_n = \sum_{i=1}^{n} X_n$ is then distribution

$$S_n \quad \overset{n\to\infty}{\approx} \quad N(n\mu, n\sigma^2)$$

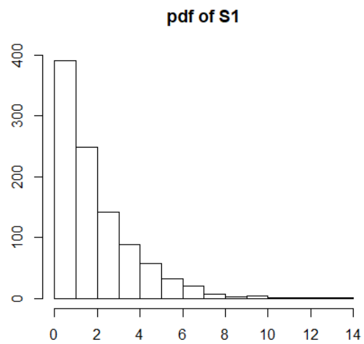
Example:

$X_i \sim Exp(0.5)$ (i.i.d.) $\to S_k = \sum_{i=1}^{k} X_i$

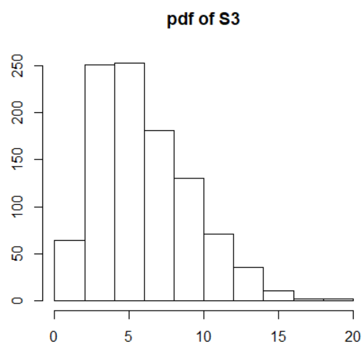
```
M <- matrix(0,50,1000)
M[1,] <- rexp(1000,lambda)
for (i in 2:50){
  M[i,] <- M[i-1,] + rexp(1000, 0.5)
}
```

```
hist(M[3,], main = 'pdf of S3', xlab='', ylab = '')
hist(M[40,], main = 'pdf of S40', xlab='', ylab = '')
```

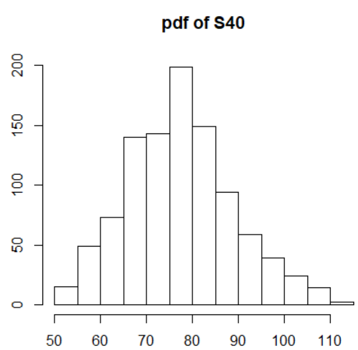
We see that as we increase the number of X considered, the random variable $S_k = \sum_{i=1}^{k} X_i$ (=the sum of the X) behaves like a normal distribution, although each X is in fact an exponential distribution.



$$S_1 = \sum_{i=1}^{1} X_i = X_1$$



$$S_3 = \sum_{i=1}^{3} X_i = X_1 + X_2 + X_3$$



$$S_{40} = \sum_{i=1}^{40} X_i$$

Note that the Central Limit Theorem also tells us something about the standard error of the sample mean $\bar{X}$:

- The standard error of $\overline{X}$ is given by $\frac{\sigma}{\sqrt{n}}$.

- In most practical situations σ is not known but rather estimated.

- The estimated standard error (SE) is:

$$\frac{s}{\sqrt{n}} = \frac{1}{\sqrt{n}} \underbrace{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}}_{s} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n(n-1)}}$$

Example:

For a temperature of 100°F and 550 watts, the following measurements of thermal conductivity were obtained:

41.60	41.48	42.34	41.95	41.86
42.18	41.72	42.26	41.81	42.04

$\rightarrow$ sample mean is $41.924 = \frac{1}{10}(41.60 + 41.48 \cdots + 42.04)$

$\rightarrow$ estimated standard error is sample standard deviation s divided by $\sqrt{10}$, here $\sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n(n-1)}} = \sqrt{\frac{(41.60^2 + 41.48^2 \cdots + 42.04^2) - 10 \cdot 41.924^2}{10 \cdot 9}} = 0.0898$

Confidence Interval

confidence interval for μ (the real mean):

$$l \leq \mu \leq u,$$

- Let $X_1, \ldots, X_n$ be collected data

- Endpoints are values of random variables $L = g_1(X_1, \ldots, X_n)$ and $U =$

$g_2(X_1, \ldots, X_n)$ such that

$$P(L(\mathbf{X}) \leq \mu \leq U(\mathbf{X})) = 1 - \alpha, \quad \alpha \in (0, 1).$$

$\longrightarrow 1 - \alpha$ is called the **confidence level**.

((l, u) is the $100 \cdot (1 - \alpha)$ % confidence interval.)

Confidence Interval for the Mean

Let X_i be i.i.d., then:

- Recall
$$\overline{X} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

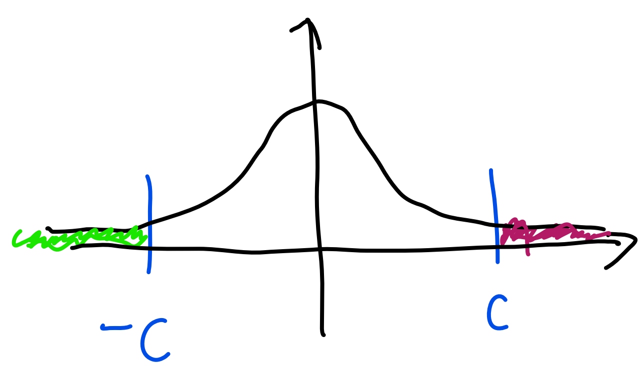
- That is, for some positive scalar value c, we have
$$P\left(\overline{X} \leq \mu + c\frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq c\right) = \Phi(c)$$
  and
$$P\left(\overline{X} \geq \mu - c\frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq -c\right) = 1 - P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq -c\right) = \Phi(-c) = 1 - \Phi(c)$$

Why is $\Phi(-c) = 1 - \Phi(c)$?

Left-hand side: $\Phi(-c) = P(Z \leq -c) \rightarrow$ green area. ($Z \sim N(0, 1)$).

Right-hand side: $1 - \Phi(c) = 1 - P(Z \leq c) = P(Z > c) \rightarrow$ purple area.

Due to symmetry of the standard normal distribution, the green and purple areas are exactly the same!



- Together, we have

$$P\left(\mu - c\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + c\frac{\sigma}{\sqrt{n}}\right) = P\left(\overline{X} \le \mu + c\frac{\sigma}{\sqrt{n}}\right) - P\left(\overline{X} \le \mu - c\frac{\sigma}{\sqrt{n}}\right)$$
$$= \Phi(c) - (1 - \Phi(c)) = 2\Phi(c) - 1.$$

- The previous is equal to

$$P\left(\mu - c\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + c\frac{\sigma}{\sqrt{n}}\right) = P\left(\overline{X} - c\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + c\frac{\sigma}{\sqrt{n}}\right)$$

- So

$$P\left(\overline{X} - c\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + c\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(c) - 1$$

- Recall that we want

$$P\left(\overbrace{\overline{X} - c\frac{\sigma}{\sqrt{n}}}^{l} \le \mu \le \overbrace{\overline{X} + c\frac{\sigma}{\sqrt{n}}}^{u}\right) = 1 - \alpha,$$

- So, we need

$$2\Phi(c) - 1 = 1 - \alpha$$

$$\Rightarrow \alpha = 2(1 - \Phi(c)) \Rightarrow \Phi(c) = 1 - \frac{\alpha}{2}$$

$\rightarrow c$ is often denoted by $z_{1-\frac{\alpha}{2}}$.

Note: $\alpha = 2(1 - \Phi(c)) = 2\Phi(-c) \rightarrow \frac{\alpha}{2} = \Phi(-c)$

Confidence Interval for Mean – Summary

The $100(1-\alpha)\%$ confidence interval on μ is

$$\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

for $\alpha = 2\Phi(-z_{1-\alpha/2})$.

1. $99\% \Rightarrow \alpha = 0.01 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.005 \Rightarrow z_{1-\alpha/2} = 2.57$

2. $98\% \Rightarrow \alpha = 0.02 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.01 \Rightarrow z_{1-\alpha/2} = 2.32$

3. $95\% \Rightarrow \alpha = 0.05 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.025 \Rightarrow z_{1-\alpha/2} = 1.96$

4. $90\% \Rightarrow \alpha = 0.1 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.05 \Rightarrow z_{1-\alpha/2} = 1.64$

Acceptable Sample Size

Since

$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

$\rightarrow$

$$P\left(\left|\overline{X} - \mu\right| \le z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

$\rightarrow$ If we pick

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\kappa}\right)^2,$$

then

$$P\left(\left|\overline{X} - \mu\right| \le \kappa\right) = 1 - \alpha,$$

$\rightarrow$ we can be $100(1-\alpha)\%$ confident that the error $|\overline{x} - \mu|$ will not exceed a specified amount κ

Example: Sample size

Let the (true) waiting time for an Uber is exponentially distributed with mean $= 2$ minutes. How many Uber-users must be questioned to ensure that the error between the same mean and true mean is at most 0.2 with a confidence of 98%?

Answer:

- $\kappa = 0.2$
- Confidence should be 98% $\to 1 - \alpha = 0.98 \to \alpha = 0.2$
- Look in the normal table to find $z_{1-\frac{\alpha}{2}}$ such that $\Phi(z_{1-\frac{\alpha}{2}}) = 0.98 \longrightarrow z_{1-\frac{\alpha}{2}} = 2.32$
- Get σ by realizing that we have exponentially distributed R.V. with mean 2, that is $E[X] = 2 = \frac{1}{\lambda}$ and we recall that the Variance is $\sigma^2 = \text{Var}(X) = \frac{1}{\lambda^2} = 4$ and therefore $\sigma = 2$

$$\Longrightarrow n = \left(\frac{z_{1-\alpha/2}\sigma}{\kappa}\right)^2 = \left(\frac{2.32 \cdot 2}{0.2}\right)^2 = 538.24$$

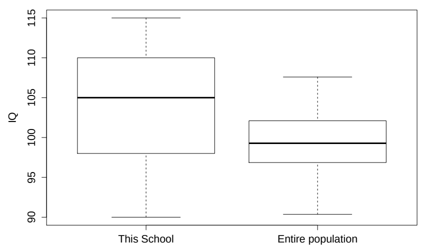
$\to$ You would have to ask at least 539 Uber-users to have at most an error of 0.2 with confidence 98%.

Hypothesis testing

Example: School choice

School A claims that its students have a higher IQ with $IQ_A \sim N(105, 25)$ and the IQ of all schools is $IQ_{all} \sim N(100, 16)$.

- What decision should we make?
- Is the observed data is due to **chance**, or,
- due to **effect**?



Example: Medical treatment

14 cancer patients were randomly assigned to either the control or treatment group.

	Survival (days)	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

- Did the treatment prolong the survival?
- Is the observed data is due to **chance**, or, due to **effect**?

Question

Is the observed *data* is due to **chance**, or due to **effect**?

$\Downarrow$

Formulate two(mutually exclusive) hypothesis:

Research Hypotheses

1. The *null* hypothesis H_0, which stands for our initial assumption about the data.
2. The *alternative hypothesis* H_1, (sometimes called H_A).

Criminal Trial:

- H_0 : Defendant is **not guilty**.
- H_1 : Defendant is **guilty**.

Choosing a school:

- H_0 : The observed IQ in the school is due to **chance**.
- H_1 : The observed IQ in the school is due to **effect**. (One should definitely prefer this school!)

Medical treatment:

- H_0 : The treatment does not prolong the survival (i.e., the observed data is due to '**chance**).
- H_1 : The treatment does prolong the survival (i.e., the observed data is due to '**effect**)

Procedure:

1. Collect the data.
2. Formulate H_0 and H_1.
3. Based on the data, decide whether to reject or not reject H_0.

Open Question: How to decide whether to reject H_0? (We will get back to that.)

	True state	
Decision	H_0 true	H_1 true
Accept H_0	OK	Type II Error (false negative)
Reject H_0	Type I Error (false positive)	OK

Statistical Test Errors

- The probability of a Type I Error is called the **significance level of the test**, denoted by α.

$$\alpha = P(\text{Type I Error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

(Common choice: $\alpha = 0.05$, i.e., accepting to have a 5% probability of incorrectly rejecting the null hypothesis.)

- The probability of a Type II Error is called the **power of the test**, denoted by β.

$$\beta = P(\text{type II error}) = P(\text{accept } H_0 \mid H_1 \text{ is true})$$

$\longrightarrow$ **AIM:** α is low and power $(1 - \beta)$ is large.

Remarks to α and β

- In most hypothesis tests used in practice (and in this course), a specified level of type I error, α is predetermined (e.g. $\alpha = 0.05$) and the type II error is not directly specified.

- The probability of making a type II error also depends on the sample size n - increasing the sample size results in a decrease in the probability of a type II error.

- The population (or natural) variability/variance also affects the power β.

Open Question

How to decide whether to reject H_0?

- Let X be a random variable with range $\mathcal{X}$.
- Find an "appropriate" subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**. Often

$$R = \{x : T(x) > c\},$$

where T is some **test statistic** and c is called a **critical value**.

- Then decide via the rule:

$$\begin{cases} X \in R \Rightarrow & \text{reject the null hypothesis } H_0 \\ X \notin R \Rightarrow & \text{do not reject the null hypothesis.} \end{cases}$$

School Example - revisited

Suppose we gathered some data $X_1, \ldots, X_n$ from this private school.

- A possible **test statistics** $T(X_1, \ldots, X_n)$ could be:

$$T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i - 100 = \overline{X} - 100,$$

since 100 is the expected value of the entire population.

- Let H_0 be the hypothesis that the higher IQ is due to chance.
- Reject H_0 if T is "large" (question: What is large?).

- Specify "large" via the **critical value** c. For example $c = 4$, then

$$R = \{x_1, \dots x_n : T(x) > c\} = \left\{x_1, \dots x_n : \bar{X} > 104\right\}$$

Finding critical value:

- Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathsf{N}(\mu, \sigma^2)$, ($\sigma$ is known).
- $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$

  $\Theta = [\mu_0, \infty), \quad , \Theta_0 = \{\mu_0\}, \quad , \Theta_1 = (\mu_0, \infty)$.
- Choose $T = \overline{X}$.
- Let $R = \left\{x_1, \dots, x_n : \overline{X} > c\right\}$.
- Let $\alpha = 0.05$.

$$0.05 = \overbrace{\alpha = P_{\mu_0}(\overline{X} > c)}^{\text{Type I error}} = P_{\mu_0}\left(\frac{(\overline{X} - \mu_0)}{\sigma/\sqrt{n}} > \frac{(c - \mu_0)}{\sigma/\sqrt{n}}\right)$$
$$\overset{CLT}{=} P\left(Z > \frac{(c - \mu_0)}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{(c - \mu_0)}{\sigma/\sqrt{n}}\right).$$

Solve for c:

$$0.05 = \alpha = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

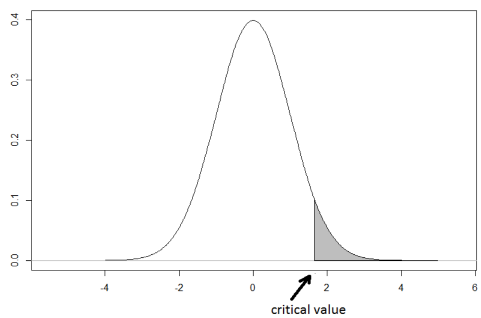
Therefore

$$0.95 = \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right)$$

Note: Since the cdf Φ is an increasing function, as c becomes larger then the left-hand side becomes also larger, which implies that α is smaller.

$$1.96 = \frac{\sqrt{n}(c - \mu_0)}{\sigma}$$

If n, μ_0 and σ are given, one can solve for c!



- The area of the shaded area is α!

- If the observed test statistics falls into the shaded area, we **reject the null hypothesis**.

Example: An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%. They abducted a random sample of 400 humans, performed special psychological tests, and found that 14% of the sample are rebellious. The true standard deviation is 0.25.

$$T = \bar{X}, \quad H_0 : \mu \leq 0.1, \quad H_1 : \mu > 0.1, \quad \alpha = 0.05$$

Idea: Find the critical value c and see if the sample mean is above.

$$0.05 = 1 - \Phi\left(\frac{(c - \mu_0)}{s/\sqrt{n}}\right) \quad \leftrightarrow \quad \Phi\left(\frac{(c - 0.1)}{0.25/\sqrt{400}}\right) = 0.95$$

$$1.65 = \frac{\sqrt{400}(c - 0.1)}{0.25} \quad \leftrightarrow \quad c = 0.1206$$

$\rightarrow$ Rejection Region $\mathcal{R} = \{T(X) = \bar{X} > c = 0.1206\}$

Since $\bar{x} = 0.14$, $\bar{x} \in \mathcal{R} \implies$ Reject $H_0 \leftrightarrow$ the "true" proportion of rebellious humans is not only 10%.