Imperial College
London

23 May 2019

# Introduction to Sampling and Hypothesis Testing

Yuan Qin

Imperial College London

# Inroduction to Sampling and Hypothesis Testing

**Yuan Qin**

PhD Candidate
Electrical and Electronic Engineering
Imperial College London

Email: yuan.qin14@imperial.ac.uk

Acknowledgments: Dr Richard Bale and Dr Fabiana Gordon

Imperial College London

# Learning Outcomes

- ☐ **Random Variables and Distributions**

- ☐ **Sampling**
  - Central Limit Theorem
  - Sampling Distribution
  - Sampling Variability

- ☐ **Statistical Inference**
  - Confidence Intervals
  - Hypothesis Testing

# Review: Descriptive Statistics

| | |
|---|---|
| Arithmetic Mean | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ |
| Median | Middle observation |
| Mode | Value occurs most frequently |
| Variance | $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ |
| Standard Deviation | $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ |

# Random Variables

☐ **Random variable**: A variable whose possible values are outcomes of a random phenomenon

☐ Discrete random variable: Countable number of random variables in a given interval

- Toss a coin, roll a dice

☐ **Continuous random variable**: Unaccountably infinite number of random variables in a given interval, could be any value

- Daily rainfall, waiting time at a traffic light
- **Probability density function $f(x)$**:

$$P(a < X < b) = \int_a^b f(x)dx$$

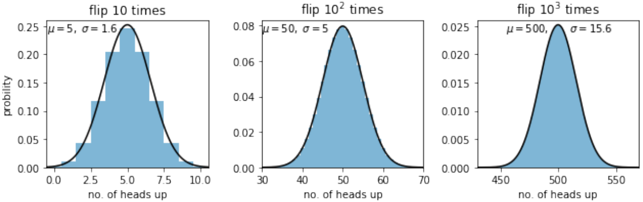
Imperial College London

# Distributions

☐ **Probability Distribution**: mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

☐ The distribution of many variables can be approximated by **Normal (Gaussian) Distribution** with known properties.

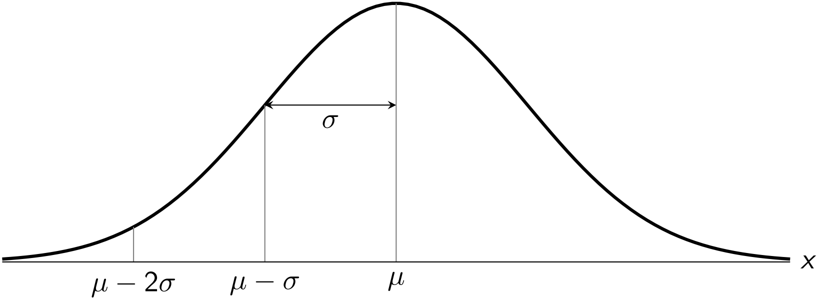
☐ Example: flip a coin 10, $10^2$, $10^3$ times, prob. of head up x times $p(x) = \binom{n}{x}\left(\frac{1}{2}\right)^n$



# The Normal Distribution

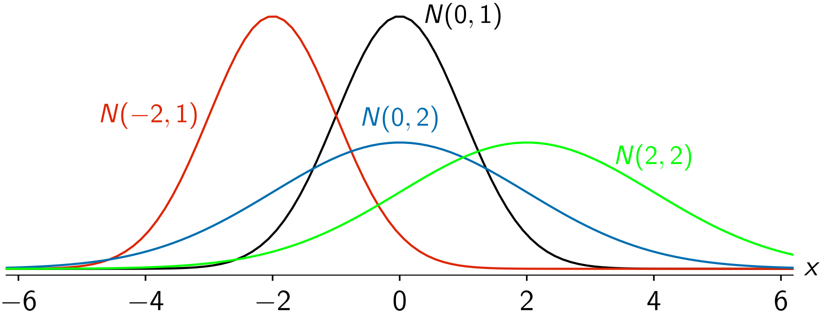
☐ The Normal distribution has the shape of a "bell curve" with parameters $\mu$ and $\sigma^2$ that determine the center and spread.

☐ Probability density function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$



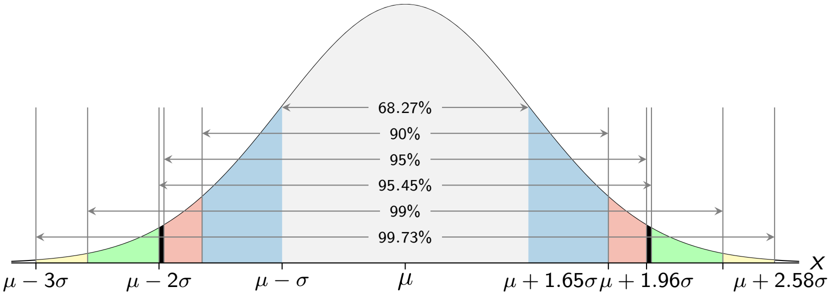
# The Normal Distribution (cont'd)

- Each different value of $\mu$ and $\sigma^2$ gives a different Normal distribution, denoted by $N(\mu, \sigma^2)$.



# Properties of the Normal Distribution

| | | | |
|---|---|---|---|
| Intervals | $(\mu - \sigma, \mu + \sigma)$ | $(\mu - 2\sigma, \mu + 2\sigma)$ | $(\mu - 3\sigma, \mu + 3\sigma)$ |
| Probability | 68.27% | 95.45% | 99.73% |
| Intervals | $\mu \pm 1.65\sigma$ | $\mu \pm 1.96\sigma$ | $\mu \pm 2.58\sigma$ |
| Probability | 90% | 95% | 99% |



# Standardization

☐ If $N(0,1)$, it is called **Standard Normal distribution**

☐ Any non-standard Normal distribution can be transformed into the standard one. This process is called **standardization**

☐ Why standardization:

$$P(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx \to P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}} dx$$

☐ **Linear transformation** is used to convert a non-standard Normal distribution into the standard one

☐ If $X \sim N(\mu, \sigma^2)$, we want $Z \sim N(0,1)$, then we use the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

# Practice 1

☐ Fill in the blanks

| $X \sim N(0,1)$ | | | | | | |
|---|---|---|---|---|---|---|
| Intervals | $(-1,1)$ | | | $(-2,2)$ | | $(-3,3)$ |
| Probability | | 90% | 95% | | 99% | |

☐ An array $\mathbf{X} = [0,1,2,3,4,5]$ follows the Normal distribution $N(7,25)$. Please convert this array and let it follow the Standard Normal distribution.

# Solution 1

- $X \sim N(7, 25)$ gives $\mu = 7$, $\sigma = 5$
- and $\mathbf{Z} = (\mathbf{X} - \mu)/\sigma$
- Hence, $\mathbf{Z} = [-1.4, -1.2, -1.0, -0.8, -0.6, -.4]$

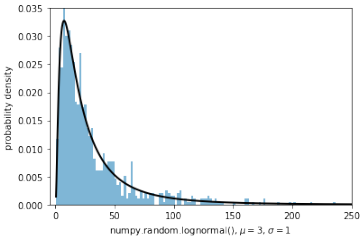
Imperial College London

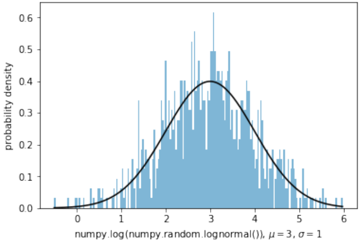
# Log Transformation for Skewed Distributions

☐ A distribution that is not symmetric but exhibit skewness with heavy right tail

☐ When appropriate the Log transformation has good statistical properties

☐ X is log-normal if log(X) is normally distributed and this property is key to transform your data back to its original scale



# Sampling

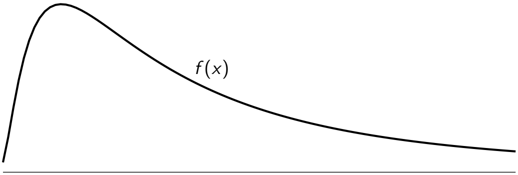
- ☐ Statistical inference: using observed data to estimate characteristics of the whole population
  - Assumed that the observed data is sampled from population
  - Characteristics: mean, variance, correlation, etc ...
- ☐ Factor that influences our ability to make inference from a sample to a population:
  - **Bias**: sample is not a representative of the whole population
- ☐ Methods to avoid bias:
  - Random sampling: each element in the population has an equal probability of selection
  - Stratified sampling: random selection in each strata
  - Systematic sampling, cluster sampling, etc ...

# Sampling Variability

☐ Example: Random sampling from Lognormal distribution $lnX \sim N(0,1)$, where $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp(\frac{-(lnx-\mu)^2}{2\sigma^2})$

☐ 5 trials, 50 samples per trial. Each trial gives a different $\overline{lnX}$ and $s$

| | $\overline{lnX}$ | $s$ |
|---|---|---|
| 1 | -0.361 | 1.065 |
| 2 | 0.205 | 0.965 |
| 3 | -0.138 | 0.914 |
| 4 | -0.100 | 0.983 |
| 5 | 0.114 | 0.996 |



# Distribution of the Sample Mean

☐ What is the distribution of sample mean $\overline{X}$, or $\overline{lnX}$ in the previous example?

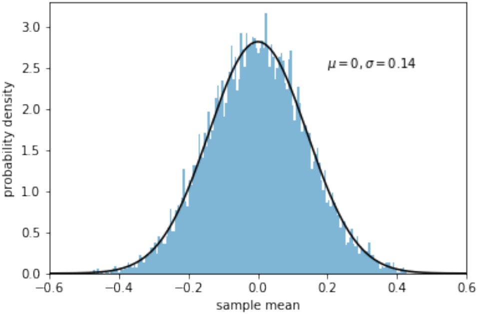
☐ **Central Limit Theorem**: If a random sample $X$ with n observations is drawn from a population with finite mean $\mu$ and variance $\sigma^2$, then, when n is sufficiently large ($n > 30$), the sample mean $\overline{X}$ can be approximated by $\overline{X} \sim N(\mu, \sigma^2/n)$

- Sufficiently large: typically $n > 30$
- Distribution of X is not necessarily Normal
- Allow us use sample mean $\overline{X}$ to make statistical inferences about the population mean $\mu$

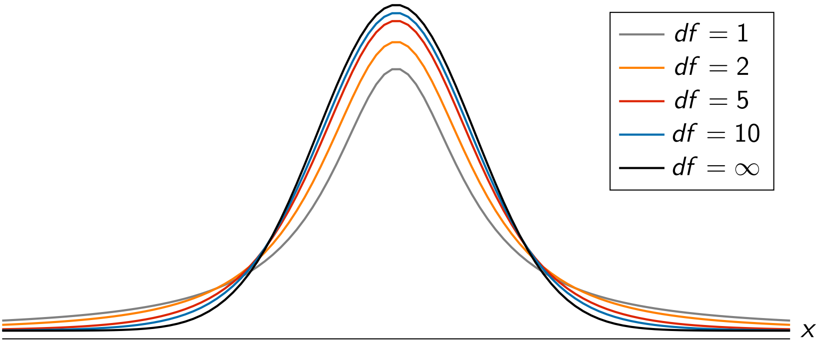
# Distribution of the Sample Mean (cont'd)

☐ Example: $lnX \sim N(0, 1)$, 10000 trials with 50 samples per trial. Plot the sample mean $\overline{ln\mathbf{X}}$

☐ According to Central Limit Theorem, we derive that the sample mean $\overline{X}$ can be approximated by $\overline{X} \sim N(0, 1/50)$



# What if $n \leq 30$? Student's t-Distribution



# Student's t-Distribution

- When $n \leq 30$ and $\sigma$ unknown, use **Student's t-distribution** instead of Normal distribution

- Closely related to standard Normal distribution
  - Continuous probability distribution
  - Heavier tails
  - 1 additional parameter: degree of freedom (df)

- For univariate t-distribution, $df = n - 1$

# Sampling Variability and Standard Error

☐ **Standard Error** (SE) of the mean is the standard deviation (SD) of the sample mean, which quantifies the accuracy of the sample mean $\overline{X}$ as an estimate of the population mean $\mu$

☐ Standard error calculation

- If $\sigma$ is known, SE is calculated by $SE = \sigma/\sqrt{n}$
- If $\sigma$ is unknown, SE is estimated by $SE \approx s/\sqrt{n}$
- Where $\sigma$ is the population standard deviation, $s$ is the sample standard deviation, and $n$ is the sampling size

# Standard Deviation vs. Standard Error

☐ Standard error: quantifies the typical error or difference between the sample mean $\overline{X}$ and the theoretical mean $\mu$ in the population from which the sample was drawn

$$SE = \frac{\sigma}{\sqrt{n}}$$

☐ Standard deviation of a sample of observations measures how a typical observation in the sample deviates from the sample mean

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

# Practice 2

☐ You are investigating the Body Mass Index for people with different diet, calculate standard deviation and standard error for the two diet groups

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Diet 1: low consumption of fruit/veg | | | | | | | | | | | |
| 24.1 | 23.5 | 18.5 | 16.7 | 26.3 | 28.5 | 25.2 | 23.4 | 22.5 | 29.9 | | |
| Diet 2: high consumption of fruit/veg | | | | | | | | | | | |
| 22.1 | 20.5 | 18.5 | 16.9 | 24.1 | 26.0 | 22.2 | 28.4 | 21.5 | 31.9 | 23.0 | 17.2 |

# Solution 2

- $n_{d1} = 10$, $n_{d2} = 12$
- $\mu_{d1} = 23.86$, $\mu_{d2} = 22.69$
- $SD_{d1} = 4.06$, $SD_{d2} = 4.47$
- $SE_{d1} = 1.28$, $SE_{d2} = 1.29$

# Confidence Intervals

☐ **Confidence interval** (CI) quantifies the level of confidence that the parameter lies in the interval

- e.g. level of confidence that CI contains the true value of the population mean $\mu$

☐ Factors required to construct a confidence interval for the true value of the population mean

- Sample mean $\overline{X}$
- Standard error of the mean $\sigma/\sqrt{n}$
- Sample mean distribution, e.g. Normal distribution
- Level of confidence, e.g. 95%

# Confidence Intervals (cont'd)

- Constructing a 95% confidence interval, when $n > 30$
  - If $X \sim N(\mu, \sigma^2)$, $\overline{X} \sim N(\mu, \sigma^2/n)$
  - Standardization $\overline{X}$ gives:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{1}$$

  - 95% confidence interval refers to $P(-1.96 < Z < 1.96) = 0.95$
  - Replace $Z$ by equation (1) gives:

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

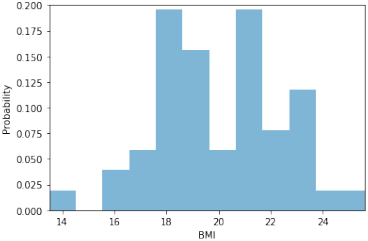
# Confidence Intervals (cont'd)

- Constructing a 95% confidence interval, when $n > 30$
  - Therefore, 95% confidence interval for the true population mean $\mu$ is
$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$
  - Out of 100 samples, if we obtain a CI as above, 95 will contain the true but unknown population mean $\mu$
  - Or, we are 95% sure that the true but unknown population mean $\mu$ lies on CI obtained from a sample mean $\overline{X}$

# Confidence Intervals: Example

- Investigating BMI for women living in rural areas of Bangladesh

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18.1 | 21.6 | 19.5 | 21.0 | 26.8 | 19.4 | 21.0 | 23.8 | 19.1 | 19.5 | 21.7 | 22.9 | 22.9 |
| 18.0 | 18.1 | 17.0 | 17.6 | 15.6 | 21.3 | 19.8 | 22.3 | 20.8 | 18.1 | 22.1 | 23.6 | 19.2 |
| 18.6 | 25.6 | 21.6 | 17.7 | 20.9 | 20.8 | 23.5 | 15.9 | 20.0 | 18.6 | 23.2 | 22.2 | 21.3 |
| 13.5 | 16.8 | 19.4 | 21.2 | 18.8 | 23.2 | 18.0 | 17.5 | 18.0 | 19.5 | 20.6 | | |




| Summary Statistics | |
|---|---|
| Mean | 20.14 |
| Standard Deviation | 2.60 |
| Sample Size | 50 |
| Standard Error | 0.37 |

# Confidence Intervals: Example (cont'd)

- Sample mean $\overline{X} = 20.14$, however it is probably not exactly equal to the population mean $\mu$
- We need to use sample mean to calculate a confidence interval of likely values for the population mean $\mu$
- Assume:
  - Each observation is independently and randomly sampled from the population with unknown mean $\mu$
  - Population standard deviation $\sigma$ is equal to the sample standard deviation $s$, i.e. $\sigma = s$
- Then, $\overline{X} \sim N(\mu, \sigma^2/n)$

# Confidence Intervals: Example (cont'd)

□ A 95% confidence interval is given by

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

□ Replacing $\sigma$ and $n$ from sample statistics, we have

$$(\square\square.\square\square, \square\square.\square\square)(\textit{practice } 3)$$

□ We are 95% sure that the true population mean $\mu$ lies within this confidence interval

# Solution 3

- ☐ $(\overline{X} - 1.96SE, \overline{X} + 1.96SE)$
- ☐ $\overline{X} = 20.14,\ SE = 0.37$
- ☐ $(19.41, 20.87)$

# Critical Value

- ☐ A 95% confidence interval $(\overline{X} - 1.96\sigma/\sqrt{n}, \overline{X} + 1.96\sigma/\sqrt{n})$
  - 95%: **confidence level**
  - 1.96: **critical value** $(Z)$

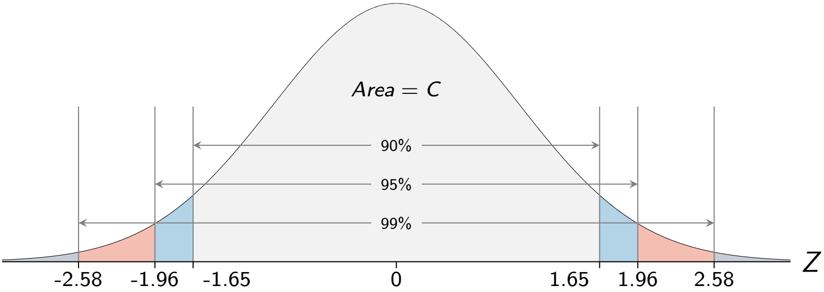
☐ General formula for a $100C\%$ confidence interval for population mean $\mu$ when $\sigma$ is known

$$(\overline{X} - Z\frac{\sigma}{\sqrt{n}},\ \overline{X} + Z\frac{\sigma}{\sqrt{n}})$$

- where $Z^*$ is the value from a standard Normal table that gives you a tail probability of $(1 - C)/2$
- A 95% confidence interval where $C = 0.95$ gives a tail probability of 0.025

# Critical Value: Example

| Confidence Level | 90% | 95% | 99% |
|---|---|---|---|
| Critical Value $Z$ | 1.65 | 1.96 | 2.58 |



# Confidence Intervals: What if $n \leq 30$?

| | $n > 30$ | $n \leq 30$ |
|---|---|---|
| Known $\sigma$ | $Z \sim N(0,1)$ | $Z \sim N(0,1)$ |
| Unknown $\sigma$ | $Z \sim N(0,1)$ (approx.) | $T \sim t_{df}$ |

- For $Z \sim N(0,1)$, we have

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

- For $T \sim t_{df}$, where $t_{df}$ is a Student t-distribution, we have

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

# Confidence Intervals: Student's t-Distribution

☐ If $\sigma$ is unknown and sample size is large ($n > 30$), $t_{df}$ where $df = n - 1$ approximates the standard Normal distribution

☐ General formula for a $100C\%$ confidence interval for population mean $\mu$ when $\sigma$ is unknown and $n \leq 30$:

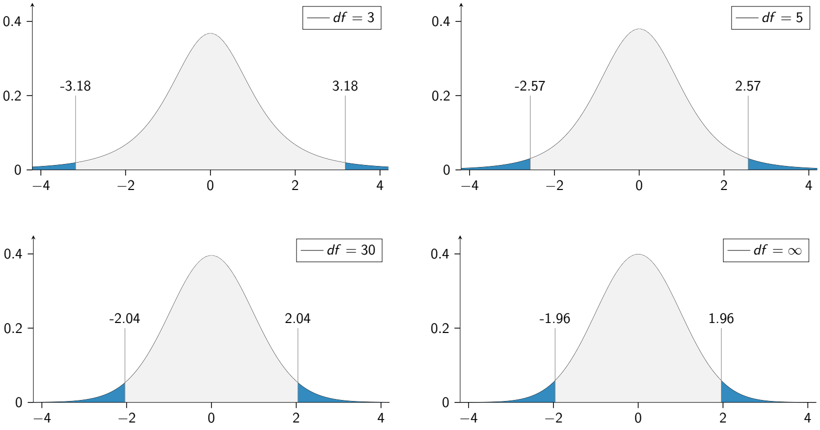
$$(\overline{X} - Ts/\sqrt{n},\ \overline{X} + Ts/\sqrt{n})$$

☐ Example: 95% confidence interval for 9 observations ($n = 9$) with known and unknown $\sigma$

| Known $\sigma$ | Unknown $\sigma$ |
|---|---|
| $(\overline{X} - 1.96\sigma/3,\ \overline{X} + 1.96\sigma/3)$ | $(\overline{X} - 2.31s/3,\ \overline{X} + 2.31s/3)$ |

Imperial College
London

# 95% Confidence Interval Tail Areas



# Confidence Intervals: Student's t-Distribution

☐ A confidence interval based on Student t-distribution will be wider than the corresponding interval based on the Normal distribution

☐ **Intuition**

- Sample standard deviation is likely to be poor estimate of population standard deviation $\sigma$, when sample size is small
- Confidence interval using sample standard deviation $s$ should be wider than using $\sigma$

☐ Student t-distribution with large sample size ($n > 30$) is approximately the standard Normal distribution, so confidence intervals will be about the same under both distributions

# Difference between Two Population Means: Example

☐ The birth weights of babies have been measured for a sample of mothers splitting into two categories: nonsmoking and heavy smoking

- Two categories are measured independently from each other
- Both come from Normal distributions
- Same unknown variance $\sigma$

☐ We want to know is there a significant difference in mean birth weights between the two categories

# Difference between Two Population Means: Example (cont'd)

☐ Summary of Data

| Category | Non-smoking | | | | Heavy smoking | | | |
|---|---|---|---|---|---|---|---|---|
| Weight (kg) | 3.99 | 3.79 | 3.60 | 3.73 | 3.18 | 2.84 | 2.90 | 3.27 |
| | 3.21 | 3.60 | 4.08 | 3.61 | 3.85 | 3.52 | 3.23 | 2.76 |
| | 3.83 | 3.31 | 4.13 | 3.26 | 3.60 | 3.75 | 3.59 | 3.63 |
| | 3.54 | 3.51 | 2.71 | | 2.38 | 2.34 | | |
| Sample size | $n_1 = 15$ | | | | $n_2 = 14$ | | | |
| Mean | $\overline{X}_1 = 3.59$ | | | | $\overline{X}_2 = 3.20$ | | | |
| Sample SD | $s_1 = 0.37$ | | | | $s_2 = 0.49$ | | | |

# Difference between Two Population Means: Example (cont'd)

- The difference between the sample means

$$d_{\overline{X}} = \overline{X}_1 - \overline{X}_2 = 0.39$$

- The difference $d_{\overline{X}}$ arises by chance, or it is significant?
- Build the confidence interval for the difference between population means $d_\mu = \mu_1 - \mu_2$
  - $(d_{\overline{X}} - t_{df}\, s_E,\ d_{\overline{X}} + t_{df}\, s_E)$
  - If CI contains 0, $d_{\overline{X}}$ could arise by chance
  - Otherwise, it could be significant

# Practice 4

- ☐ Build the confidence interval for $\mu$ at 95% confidence level, given following equations derived from the Student's t-distribution:
  - Sample SE for difference of two sample means:

$$s_E^2 = s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

  - Where

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

  - Critical value at $df = 27$: $t_{df} = 2.05$
- ☐ What conclusion can we draw? [ans: $(0.06, 0.72)$]

# Hypothesis Testing

- ☐ Alternative method: other than constructing a confidence interval, we assume that there is no difference in mean birth weights between the two categories, i.e. $d_\mu = 0$. Then we calculate under which condition the assumption does not hold
- ☐ The alternative method is called the **hypothesis testing**
- ☐ Hypothesis
  - **Null hypothesis** ($H_0$): A default position that there is no effect or change between two measured phenomena, any differences are due to chance. e.g. $d_\mu = 0$
  - **Alternative hypothesis** ($H_1$): the counterpart of $H_0$. e.g. $d_\mu \neq 0$

# Hypothesis Testing: Difference between Two Population Means (cont'd)

☐ $H_0$: there is no difference in mean birth weights between the two categories, $d_\mu = 0$

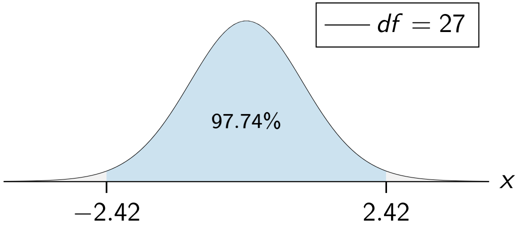
☐ $H_1$: there is difference in mean birth weights between the two categories, $d_\mu \neq 0$

☐ Difference in sample means $d_{\overline{X}} = 0.39$ is a random variable of Student's t-distribution with $df = 27$, standardize $d_{\overline{X}}$ by

$$t_{df} = \frac{d_{\overline{X}} - d_\mu}{s_E}$$

# Hypothesis Testing: Difference between Two Population Means (cont'd)

- For $X \sim t_{df=27}$, $t_{df} = 2.42$, $P(|X| < t_{df}) = 97.74\%$
- We are 97.74% sure to say that men birth weights between the two categories are not the same (reject $H_0$)
- Implies there is difference in mean birth weights between the two categories ($H_1$)



# Hypothesis Testing Terminologies

☐ $H_0$ and $H_1$ are always two **rival** hypothesis, e.g. $H_0$: $\mu = 0$, $H_1$: $\mu \neq 0$

☐ **Test statistic**: A quantity derived from sample used in hypothesis testing, e.g. $t_{df}$

☐ **P-value**: The probability that the observed data is larger than test statistic, given the null hypothesis is true

- e.g. $p = 0.03$
- The smaller the p-value is, the more likely to reject $H_0$

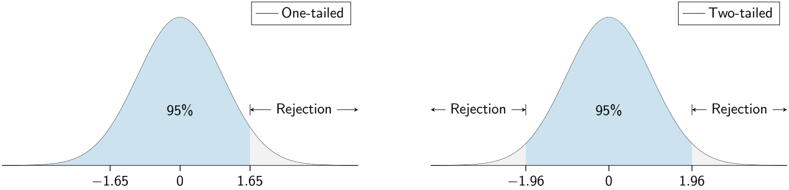
☐ **Significance level** $\alpha$ is the probability that p-value under which $H_0$ will be rejected

- e.g. If $\alpha = 0.05$, $p = 0.03 < \alpha$, reject $H_0$

Imperial College London

# One-Tailed Test vs Two-Tailed Test

- Two-tailed test: $H_0$: $\mu = 0$, $H_1$: $\mu \neq 0$
  - For z-test with a given significance level $\alpha = 0.05$, $H_0$ is rejected when $P(|X| > z_{\alpha/2}) < \alpha/2$, where $z_{\alpha/2} = 1.96$
- One-tailed test: $H_0$: $\mu < 0$, $H_1$: $\mu \geq 0$
  - For z-test with a given significance level $\alpha = 0.05$, $H_0$ is rejected when $P(X > z_{\alpha}) < \alpha$, where $z_{\alpha} = 1.65$



# Test Statistic

☐ Measures the difference between the observed data and the null hypothesis

☐ One value, e.g. $t_{df\,27}$, that can be used to perform the hypothesis test

☐ Assumed underlying distributions have different test statistics

| | | |
|---|---|---|
| Student's t-distribution | t-test | $t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ |
| Normal distribution | z-test | $z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ |
| Chi-squared distribution | $\chi^2$ test | $\chi^2 = \sum (f_o - f_e)^2 / f_e$ |
| F-distribution | $F$-test | $F = s_1^2 / s_2^2$ |

# Hypothesis Testing Process

1. Propose a research question regarding observed data set
2. Formulate null hypothesis $H_0$
3. Calculate the test statistic
4. Find the p-value
5. Compare p-value with required significance level $\alpha$
6. Make a decision about whether to reject $H_0$

# Practice 5

☐ Given the amount of money invested (in dollars) in a plant divided by the delivered amount of energy (in quadrillion British thermal units) listed below, do these data provide sufficient evidence at $\alpha = 0.05$ to indicate a difference in the average investment/quad between gas plants and electric plants?

| **Electric** | | | | **Gas** | | | |
|---|---|---|---|---|---|---|---|
| 204.15 | 0.57 | 62.76 | 0.78 | 0.78 | 16.66 | 74.94 | 0.01 |
| 85.46 | 0.35 | 89.72 | 0.65 | 0.54 | 23.59 | 88.79 | 0.64 |
| 44.38 | 9.28 | 78.60 | | 0.82 | 91.84 | 66.64 | 7.20 |
| | | | | 0.74 | 64.67 | 165.60 | 0.36 |

# Solution 5

| | Electric | Gas |
|---|---|---|
| Sample Size, $n$ | $n_1 = 11$ | $n_2 = 16$ |
| Mean, $\overline{x}$ | $\overline{x}_1 = 52.43$ | $\overline{x}_2 = 37.74$ |
| Sample SD, $s$ | $s_1 = 62.43$ | $s_2 = 49.05$ |
| $H_0$ | $D_0 = 0$, where $D_0 = \mu_1 - \mu_2$ | |
| $H_1$ | $D_0 \neq 0$, where $D_0 = \mu_1 - \mu_2$ | |
| | $D_{\overline{x}} = \overline{x_1} - \overline{x_2} = 14.69$ | |
| df | $df = n_1 + n_2 - 2$ | |
| | $s_p^2 = 3002.05$, | $s_E = 21.46$ |
| Test Statistic | $t = (D_{\overline{x}} - D_0)/s_E = .68$ | |
| Results | $p = .5 \rightarrow p > \alpha$, $H_0$ is not rejected | |

# Type I & Type II Errors

| Rejection Table | | |
|---|---|---|
| | $H_0$ True | $H_0$ False |
| Reject $H_0$ | *Type I Error* ($\alpha$) | *Correct Decision* ($1-\beta$) |
| Fail to Reject $H_0$ | *Correct Decision* | *Type II Error* ($\beta$) |

- Two possible errors can be made when using p-values to make a decision
  - **Type I error**: reject the null hypothesis when it is true
  - **Type II error**: not reject the null hypothesis when it is false

# Type I & Type II Errors (cont'd)

- Probability of Type I and Type II errors:
  - $\alpha$: the significance level $\alpha$ is the probability of Type I error.
  - $\beta$: the probability of Type II error relative to the $H_1$ is called $\beta$
- Power of a test $1 - \beta$: probability that $H_0$ is rejected when $H_1$ is true
- Methods to reduce errors:
  - $\alpha \downarrow \longrightarrow \beta \uparrow$
  - $\beta \downarrow \longrightarrow \alpha \uparrow$
  - Increase the sample size $n$

# Example: Chi-Squared Test

- Chi-squared $\chi^2$ test
  - Test for association between two categorical variables
  - Compare the observed frequencies with the frequencies that would be expected under $H_0$
- Observations: hospital patients with and without lung cancer were asked about previous smoking habits
- Question: whether there is association between lung cancer and smoking, given $\alpha = .05$?
- $H_0$: There is no association between lung cancer and smoking
- $H_1$: Lung cancer is associated to smoking

# Observations

| | Lung Cancer (L) | Other Diseases (O) | Sum |
|---|---|---|---|
| Smoker (S) | $f_{o_{SL}} = 647$ | $f_{o_{SO}} = 621$ | 1268 |
| Non-smoker (N) | $f_{o_{NL}} = 2$ | $f_{o_{NO}} = 28$ | 30 |
| Sum | 649 | 649 | 1298 |

- $H_0$ assumes smoking and lung cancer are **independent**
- Then, $P(S, L) = P(S) * P(L)$ ...

# Expected Observations under $H_0$

| | Lung Cancer (L) | Other Diseases (O) | Sum |
|---|---|---|---|
| Smoker (S) | $f_{e_{SL}}$ | $f_{e_{SO}}$ | 1268 |
| Non-smoker (N) | $f_{e_{NL}}$ | $f_{e_{NO}}$ | 30 |
| Sum | 649 | 649 | 1298 |

- $f_{e_{SL}} = f_o P(S, L) = f_o P(S) P(L) = 1298 * \frac{1268}{1298} * \frac{649}{1298} = 634$
- $f_{e_{SO}} = 634$
- $f_{e_{NL}} = 15$
- $f_{e_{NO}} = 15$

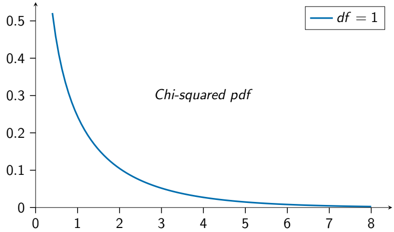
Imperial College London

# Chi-Squared Test

- The Chi-square test compares the observed frequencies ($f_{o_*}$) with the frequencies that would be expected under $H_0$ ($f_{e_*}$)
- Test statistic

$$\chi^2 = \sum \frac{(f_{o_*} - f_{e_*})^2}{f_{e_*}}$$

- $df = (no.\ of\ \ rows - 1)(no.\ of\ \ cols - 1)$



# Chi-Squared Test (cont'd)

- Computation result
  - $\chi^2 = 23.1$
  - $df = 1$
  - $p = 1.53e^{-6}$
  - $p < \alpha$
- Conclusions
  - $H_0$ is rejected due to very small p-value
  - Smoking and lung cancer are associated

# Example: $F$ - Test

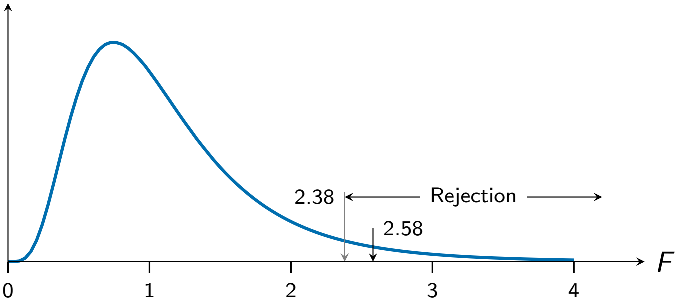
- ☐ $F$ - test: Test for the ratio of two population variances
- ☐ Observations: A company imports products from two suppliers, the weight (in kg) slightly varies for each product
- ☐ Question: Whether there is a difference between the standard deviation of product weights from two suppliers, given $\alpha = .1$?
- ☐ $H_0$: There is no difference in the standard deviation of product weights from two suppliers, i.e. $\sigma_1^2 = \sigma_2^2$
- ☐ $H_1$: The standard deviation of product weights from two suppliers are different, i.e. $\sigma_1^2 \neq \sigma_2^2$

# Summary of Observed Data

| | Supplier 1 | Supplier 2 |
|---|---|---|
| Sample Size, $n$ | $n_1 = 13$ | $n_2 = 18$ |
| Mean | 5.60 | 5.90 |
| Sample SD, $s$ | $s_1 = 3.10$ | $s_2 = 1.93$ |
| $H_0$ | $\sigma_1^2 = \sigma_2^2$ | |
| $H_1$ | $\sigma_1^2 \neq \sigma_2^2$ | |
| Test Statistic | $F = s_1^2/s_2^2$ | |
| $df$ | $df_1 = 12$ | $df_2 = 17$ |

# $F$ - Test

- $F = 2.58 \implies p = 0.04 < \alpha/2 \implies H_0$ is rejected
- The standard deviation of product weights from two suppliers are different



Imperial College London

# References


William Mendenhall, Terry Sincich. *Statistics for Engineering and the Sciences*. Pearson/Prentice Hall, Upper Saddle River, New Jersey, 2007.

Douglas G. Altman. *Practical Statistics for Medical Research*. CRC press, 1990.