

# Probability and Statistics II

Derek Li

## Contents

<b>1</b>	<b>Review of Probability</b>	<b>4</b>
1.1	Probability . . . . .	4
1.2	Expectation . . . . .	4
1.3	Indicator function . . . . .	4
1.4	Law of large number (LLN) . . . . .	5
1.5	Central limit theorem (CLT) . . . . .	5
1.6	Linear combination of Normal variables . . . . .	5
1.7	$Z$ and $\chi^2$ distribution . . . . .	6
1.8	$t$ and $F$ distribution . . . . .	6
<b>2</b>	<b>Data Collection</b>	<b>7</b>
2.1	Population and sample . . . . .	7
2.2	Parameter and statistic . . . . .	7
2.3	Finite populations . . . . .	7
2.4	Infinite populations . . . . .	8
2.5	Simple random sampling . . . . .	8
2.6	Empirical CDF . . . . .	8
2.7	Density histogram . . . . .	9
2.8	Quantile/Percentile for population . . . . .	9
2.9	Boxplot . . . . .	10
2.10	Choice of summary measures . . . . .	10
<b>3</b>	<b>Point Estimation</b>	<b>11</b>
3.1	Type of inference . . . . .	11
3.2	Method of moments estimation . . . . .	11
3.3	Maximum likelihood estimation . . . . .	12
3.4	Sampling distribution of an estimator . . . . .	13

3.5	Measuring quality of an estimator . . . . .	14
3.6	Unbiasedness . . . . .	15
<b>4</b>	<b>Sampling Distribution of <math>S^2</math></b>	<b>16</b>
4.1	Sample variance ( $S^2$ ) . . . . .	16
4.2	Sampling distribution of $S^2$ under Normal distribution . . . .	17
4.3	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$ . . . . .	18
4.4	$\chi^2_{(m)}$ . . . . .	19
<b>5</b>	<b>Properties of an Estimator: Consistency, Efficiency and Sufficiency</b>	<b>20</b>
<b>6</b>	<b>Interval Estimation</b>	<b>21</b>
6.1	Confidence interval . . . . .	21
6.2	CI for parameters of Normal distribution . . . . .	21
6.2.1	CI for $\mu$ with $\sigma^2$ known . . . . .	21
6.2.2	CI for $\mu$ with $\sigma^2$ unknown . . . . .	21
6.2.3	CI for $\sigma^2$ . . . . .	22
6.3	CI for mean of a non-Normal distribution using CLT . . . . .	22
6.4	Interpreting CI . . . . .	22
<b>7</b>	<b>Test of Hypothesis</b>	<b>24</b>
7.1	Types of hypothesis . . . . .	24
7.2	Two approaches of hypothesis testing . . . . .	24
7.2.1	Critical region approach . . . . .	24
7.2.2	$p$ -value approach . . . . .	25
7.3	Type-1, 2 error and power of a test . . . . .	26
7.4	Test of hypothesis using CI . . . . .	27
<b>8</b>	<b>Likelihood Ratio Test and Comparing Two Populations</b>	<b>28</b>
8.1	Likelihood ratio test (LRT) . . . . .	28
8.2	Constructing CI using LRT . . . . .	29
8.3	Comparing two independent Normal population . . . . .	30
8.3.1	Equality of two variances . . . . .	30
8.3.2	Equality of two means with variances known . . . . .	30
8.3.3	Equality of two means with variances unknown . . . . .	31
8.4	Comparing two population means (paired data) . . . . .	31
8.5	Comparing two populations using LRT . . . . .	31

8.6	Numerical example . . . . .	32
<b>9</b>	<b>Model Checking</b>	<b>34</b>
9.1	$\chi^2$ goodness of fit test . . . . .	34
9.2	Discrepancy statistic . . . . .	37
9.3	Residual and quantile/probability plots . . . . .	38
<b>10</b>	<b><math>\chi^2</math> Test of Independence and Homogeneity</b>	<b>41</b>
10.1	Relationship among variables . . . . .	41
10.2	Relationship of two categorical variables . . . . .	41
10.2.1	$\chi^2$ test of independence ( $X$ and $Y$ are random) . . . .	41
10.2.2	$\chi^2$ test of homogeneity ( $X$ is deterministic) . . . . .	43
<b>11</b>	<b>Correlation Coefficient and Least Square Regression</b>	<b>45</b>
11.1	Relation among quantitative variables . . . . .	45
11.2	Least square regression . . . . .	45
11.3	Classical linear regression under Normal distribution . . . . .	46
11.3.1	Properties of estimators of regression parameters . . . .	47
11.3.2	Confidence interval and $t$ -test for $\beta_2$ . . . . .	48
11.3.3	Sum of squares decomposition and ANOVA test . . . . .	49
11.3.4	Prediction and residual check . . . . .	49
11.4	Quantitative $Y$ and categorical $X$ . . . . .	50

# 1 Review of Probability

## 1.1 Probability

- The probability measure  $P$  for each event  $A$  defined on sample space  $\Omega$  satisfies the following properties:
  - $P(A)$  is non-negative and  $0 \leq P(A) \leq 1$ .
  - $P(A) = 0$  when  $A$  is empty.
  - $P(A) = 1$  when  $A$  is the entire sample space  $\Omega$ .
  - $P$  is countably additive.

## 1.2 Expectation

- Expected value/mean/average of r.v.  $X$  is defined as
  - $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) \, dx$ , when  $X$  is continuous;
  - $\mathbb{E}[X] = \sum_i x_i P(X = x_i)$ , when  $X$  is discrete.
- Expectation is a **linear operator**: Let  $X$  and  $Y$  are two r.v.s., then  $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$ .

## 1.3 Indicator function

- If  $A$  is any event, define the **indicator function** of  $A$ ,  $I_A$  to be the r.v. for all  $s \in \Omega$ ,

$$I_A(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A \end{cases}.$$

**Example 1.1.** We are rolling a dice and  $A = \{2, 4, 6\}$ .

$X$	1	2	3	4	5	6
$I_A$	0	1	0	1	0	1

Therefore,  $\mathbb{E}[I_A] = \frac{1}{6}(0 + 1 + 0 + 1 + 0 + 1) = \frac{1}{2} = P(A)$ .

## 1.4 Law of large number (LLN)

- Let  $X_1, X_2, \dots, X_i$  be a sequence of independent r.v.s. with  $\mathbb{E}[X_i] = \mu$ . Let  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then  $\bar{X}_n \xrightarrow{P} \mu$  as  $n \rightarrow \infty$ , i.e.,

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

◦ In naive words: Sample mean approaches the population mean as the sample size increases.

## 1.5 Central limit theorem (CLT)

- Suppose  $X_1, X_2, \dots$  is an i.i.d. sequence of r.v.s. each having finite mean  $\mu$  and finite variance  $\sigma^2$ . Let  $\bar{X}_n = \frac{1}{n}$ , then as  $n \rightarrow \infty$ ,  $\bar{X}_n \xrightarrow{D} \mathcal{N}(\mu, \frac{\sigma^2}{n})$  or

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} \mathcal{N}(0, 1).$$

◦ In naive words: A r.v. can follow some distribution with mean  $\mu$  and variance  $\sigma^2$ . If we pick a fixed number of samples  $n$  and calculate the sample mean repeatedly, then those sample means will have a Normal distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

## 1.6 Linear combination of Normal variables

- Let  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$  where  $i = 1, 2, \dots, n$ . Let  $Y$  be a linear combination of all the  $X_i$ 's with

$$Y = a_1 X_1 + \dots + a_n X_n + b = \sum_{i=1}^n a_i X_i + b,$$

where  $a_i, b \in \mathbb{R}$ . Then  $Y \sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i + b, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$ .

**Example 1.2.** Let  $X_1 \sim \mathcal{N}(10, 2)$ ,  $X_2 \sim \mathcal{N}(20, 3)$ ,  $Y = 0.4X_1 + 0.6X_2$ . Then  $Y \sim \mathcal{N}(16, 1.4)$ .

## 1.7 $Z$ and $\chi^2$ distribution

- Standard normal/ $\mathcal{N}(0, 1)$ / $Z$  distribution: If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ .
- $\chi^2$  distribution: Let  $U = Z^2$ , then  $U \sim \chi^2_{(1)}$ .
  - Additive property: If  $X \sim \chi^2_{(m)}$ ,  $Y \sim \chi^2_{(n)}$ , then  $X + Y \sim \chi^2_{(m+n)}$ .
  - If  $X \sim \chi^2_{(m)}$ , then  $\mathbb{E}[X] = m$ .

## 1.8 $t$ and $F$ distribution

- $t$  distribution: Let  $Z \sim \mathcal{N}(0, 1)$  and  $U \sim \chi^2_{(m)}$  be independent, then  $\frac{Z}{\sqrt{U/m}} \sim t_{(m)}$ .
- $F$  distribution: Let  $X \sim \chi^2_{(m)}$ ,  $Y \sim \chi^2_{(n)}$  be independent, then  $\frac{X/m}{Y/n} \sim F_{(m,n)}$ .

## 2 Data Collection

### 2.1 Population and sample

- **Population** is a collection of all the subjects that have something in common.
- **Sample** is a subset of the population.
  - We use the sample to make inference about the unknown characteristics of our population.
  - The sample should be representative.

### 2.2 Parameter and statistic

- **Parameter** is a characteristic (summary) of the population. For example, mean ( $\mu$ ), standard deviation ( $\sigma$ ), etc.
  - We use  $\theta$  to represent the parameter(s) of population. For example,  $X \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\theta$  stands for both  $\mu$  and  $\sigma$ .
- **Statistic** is any summary of the sample. For example, sample total ( $\sum X_i$ ), etc.
  - When a statistic is used to estimate a parameter, it is called an estimator. For example,  $S$  is an estimator of  $\sigma$ .
  - $T(X)$  is used to represent a statistic/estimator. For example, if we are dealing with sample mean, then  $T(X) = \bar{X}$ .
  - When we have observed a sample and calculate the value of an estimator, then that numerical value is called the estimate and we use lowercase letters to represent.

Parameter ( $\theta$ )	Estimator ( $T$ )	Estimate ( $t$ )
$\mu$	$\bar{X}$	$\bar{x}$
Unknown constant	Random variable	Known constant

### 2.3 Finite populations

- Let  $\pi$  represent individual subjects in a finite population  $\Pi$ . For each  $\pi$ , we have a real valued quantity  $X(\pi)$ .

- The **population CDF**,

$$F_X(x) = \frac{|\{\pi | X(\pi) \leq x\}|}{N},$$

where  $N = |\Pi|$ . Or,

$$F_X(x) = \frac{1}{N} \sum I_{(-\infty, x]}(X(\pi)) = \mathbb{E}[I_{(-\infty, x]}(X(\pi))].$$

◦ In naive words:  $F_X(x)$  is the proportion of elements in the population with their  $X$  measurement less or equal to  $x$ .

## 2.4 Infinite populations

- We use probability distributions to represent the population. Informally, we can think it as a limiting distribution of a finite population of size  $N$  when  $N \rightarrow \infty$ .

## 2.5 Simple random sampling

- With replacement:
  - Every subject of the population will have the same probability  $\frac{1}{N}$  of being selected in the sample in each draw.
  - Samples are independent.
- Without replacement:
  - Not independent.
  - If  $N \rightarrow \infty, n \ll N$ , where  $n$  is the sample size:  $P(B) = \frac{1}{N}, P(B|A) = \frac{1}{N-1}$ . But for a large  $N$  and  $n \ll N$ ,  $P(B) \approx P(B|A)$ , then samples are independent.

## 2.6 Empirical CDF

- Suppose we select a sample  $\{\pi_1, \dots, \pi_n\} \subset \Pi$ , we can approximate the population CDF  $F_X$  by the **empirical CDF**

$$\hat{F}_X(x) = \frac{|\{\pi_i | X(\pi_i) \leq x, i = 1, \dots, n\}|}{n} = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X(\pi_i)).$$



- Assuming independence, then by LLN,

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X(\pi_i)) &\xrightarrow{P} \mathbb{E}[I_{(-\infty, x]}(X(\pi_i))] = P(I_{(-\infty, x]}(X(\pi_i))) \\ &= P(X(\pi_i) \leq x) = F_X(x).\end{aligned}$$

## 2.7 Density histogram

- Suppose we have continuous variable  $X$  and can group  $X$  into intervals given by  $(h_1, h_2], \dots, (h_{m-1}, h_m]$ . The **density histogram function**

$$h_X(x) = \begin{cases} \frac{|\{\pi | X(\pi) \in (h_i, h_{i+1}]\}|}{N(h_{i+1} - h_i)}, & x \in (h_i, h_{i+1}] \\ 0, & \text{otherwise} \end{cases}.$$

◦ In naive words: In density histogram, the height of each of the bar is the relative frequency, divided by the corresponding length of the interval.

◦ When the interval lengths  $(h_{i+1} - h_i)$  gets smaller and  $N$  gets bigger, we get a smooth function.

## 2.8 Quantile/Percentile for population

- For  $p \in [0, 1]$ , the  $p$ th quantile (100 $p$ th percentile)  $x_p$ , for the distribution with CDF  $F_X$ , is defined to be the **smallest number**  $x_p$  satisfying  $p \leq F_X(x_p)$ .
  - When  $F_X$  is strictly increasing and continuous,  $x_p$  satisfies  $F_X(x_p) = p$ .
  - When  $X$  is discrete,  $F_X(x_p) = p$  may not have a solution.
- Estimating quantiles: Suppose the sample is  $(x_1, \dots, x_n)$  and after ordering we have  $x_{(1)} < \dots < x_{(n)}$ ,  $x_{(i)}$  is the  $(\frac{i}{n})$ th quantile of the empirical distribution because  $\hat{F}_X(x_{(i)}) = \frac{i}{n}$ . The sample  $p$ th quantile is  $x_p$  whenever  $\frac{i-1}{n} < p \leq \frac{i}{n}$ .
  - Linear interpolation:  $\tilde{x}_p = x_{(i-1)} + n(x_{(i)} - x_{(i-1)})(p - \frac{i-1}{n})$ .

*Proof.* We have  $\frac{\tilde{x}_p - x_{(i-1)}}{np - (i-1)} = \frac{x_{(i)} - x_{(i-1)}}{i - (i-1)}$ .

Therefore,  $\tilde{x}_p = x_{(i-1)} + n(x_{(i)} - x_{(i-1)})(p - \frac{i-1}{n})$ . □

**Example 2.1.** -2.1 -0.3 0.4 1.2 1.5 2.1 2.2 3.3 4.0 5.0

First quantile =  $Q_1 = \tilde{x}_{0.25} = x_{(2)} + 10(x_{(3)} - x_{(2)})(0.25 - \frac{2}{10}) = 0.05$

Third quantile =  $Q_3 = \tilde{x}_{0.75} = x_{(7)} + 10(x_{(8)} - x_{(7)})(0.75 - \frac{7}{10}) = 2.75$

Inter quantile range =  $IQR = Q_3 - Q_1 = 2.7$

- Median/Second quantile: We can use linear interpolation formula or

$$Q_2 = \tilde{x}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & n \text{ is even} \end{cases}.$$

## 2.9 Boxplot

- Draw a box using  $Q_1$  and  $Q_3$  as the sides and  $Q_2$  as a line inside the box.
- Lower limit =  $Q_1 - 1.5 \cdot IQR$ , Upper limit =  $Q_3 + 1.5 \cdot IQR$ .
- **Adjacent values** are the *two extreme data points* that falls within the lower and upper limit.
- **Whiskers** are the vertical lines from the quantiles to the adjacent values.
- Values beyond the adjacent values are plotted with \* and called outliers.
- If the variable is categorical, we use **bar charts**. Categories on  $x$ -axis and proportions on  $y$ -axis.

## 2.10 Choice of summary measures

- Choice of summary measures based on the skewness of the distribution
  - Mean and s.d. when distribution is symmetric.
  - Median and  $IQR$  when distribution is skewed.

## 3 Point Estimation

### 3.1 Type of inference

- Estimation:
  - Point estimation: Based on the sample observations, calculating a particular value as an estimate of the parameter.
  - Interval estimation: Calculating a range of values that is likely to contain  $\theta$ .
- Hypothesis testing: Based on the sample, assess whether a hypothetical value  $\theta_0$  is a plausible value of the  $\theta$  or not.

### 3.2 Method of moments estimation

- Let  $X_1, \dots, X_n$  be i.i.d. r.v.s. and let the  $k$ th **population moment**  $\mu_k = \mathbb{E}[X^k]$ ,  $k$ th **sample moment**  $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ .
- We use  $\hat{\mu}_k$  as an estimator of  $\mu_k$ .

**Example 3.1.**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ . Find the method of moments estimator of  $\lambda$ .

*Solution.* We have  $\lambda = \mathbb{E}[X] = \mu$ , then  $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ .

**Example 3.2.**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ . Find the method of moments estimator of  $\mu$  and  $\sigma^2$ .

*Solution.* We have  $\mu = \mathbb{E}[X]$ ,  $\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$  and thus

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{1}{n} n (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- Summary of method:
  - Express the lower order population moment(s) in terms of the parameter(s).
  - Invert the expression(s) to express the parameter(s) in terms of the population moment(s).
  - Replace the population moment(s) using the sample moment(s).

### 3.3 Maximum likelihood estimation

- Suppose  $X_1, \dots, X_n$  has a joint density or mass function  $f(x_1, \dots, x_n|\theta)$  and we observe sample  $X_1 = x_1, \dots, X_n = x_n$ . The **likelihood function** of  $\theta$ ,  $L(\theta) = f(x_1, \dots, x_n|\theta)$ .
  - If  $X$  follows a discrete distribution, it gives the **probability of observing the sample** as a function of  $\theta$ .
- If  $X_1, \dots, X_n$  are i.i.d. then  $L(\theta) = \prod_{i=1}^n f_{\theta}(x_i)$ .
  - $L(\theta)$  is not a PDF or PMF of  $\theta$ .
  - Likelihood introduces a belief ordering on parameter space  $\Omega$ . If  $L(\theta_1) > L(\theta_2)$ , the data is more likely to come from  $f_{\theta_1}$  than  $f_{\theta_2}$ .
  - The value  $L(\theta)$  is very small for every value of  $\theta$ , so often we are interested in the **likelihood ratio**  $\frac{L(\theta_1)}{L(\theta_2)}$ .
- Maximum likelihood estimation (MLE): If we are interested in a point estimation of  $\theta$ , a sensible choice will be to pick  $\hat{\theta}$  that maximizes  $L(\theta)$ , i.e.,  $L(\hat{\theta}) \geq L(\theta), \forall \theta \in \Omega$ .
  - Computation for MLE:

\* **Log-Likelihood function**

$$l(\theta) = \ln(L(\theta)) = \ln \left( \prod_{i=1}^n f_{\theta}(x_i) \right) = \sum_{i=1}^n \ln(f_{\theta}(x_i)).$$

Since  $\ln x$  is an injective increasing function of  $x > 0$ , then  $L(\hat{\theta}) \geq L(\theta), \forall \theta \in \Omega$  iff  $l(\hat{\theta}) \geq l(\theta)$ .

\* Solve  $\frac{\partial l(\theta)}{\partial \theta} = 0$  and  $\hat{\theta}$  is the solution.

\* Check if  $\left. \frac{\partial^2 l(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} < 0$ .

**Example 3.3.**  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ . Find the MLE of  $\lambda$ .

*Solution.* We have  $f(x) = \frac{e^{-\lambda} \lambda^x}{x!}$  and thus

$$L(\lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

Therefore,  $l(\lambda) = -n\lambda + \ln \lambda \sum_{i=1}^n x_i + C$ . Let  $\frac{\partial l(\lambda)}{\partial \lambda} = 0$ , we have  $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ .

◦ Properties of MLE:

\* MLE is not unique.

\* MLE may not exist.

\* The likelihood may not always be differentiable. For example,  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta], \hat{\theta} = \max\{x_1, \dots, x_n\}$ .

\* Invariance property of MLE: Let  $\hat{\theta}$  be the MLE of  $\theta$  and  $\psi(\theta)$  be any injective function of  $\theta$  defined on  $\Omega$ , then  $\psi(\hat{\theta})$  is the MLE of  $\psi(\theta)$ .

### 3.4 Sampling distribution of an estimator

- An estimator ( $T$ ) is a r.v. and if we repeat the sampling procedure and keep calculating  $T$  for each set of sample and finally draw a density histogram based on the  $T$  values, we get the sampling distribution of  $T$ .
- Assume  $X_1, \dots, X_n$  is an i.i.d. sequence of r.v.s., each having finite mean  $\mu$  and finite variance  $\sigma^2$ , then

$$\begin{aligned} \mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}X_1 + \dots + \frac{1}{n}X_n\right] = \frac{1}{n}\mathbb{E}[X_1] + \dots + \frac{1}{n}\mathbb{E}[X_n] \\ &= \frac{1}{n}n\mu = \mu, \end{aligned}$$

and

$$\begin{aligned}\text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right] = \text{Var}\left[\frac{1}{n}X_1\right] + \cdots + \text{Var}\left[\frac{1}{n}X_n\right] \\ &= \frac{1}{n^2}\text{Var}[X_1] + \cdots + \frac{1}{n^2}\text{Var}[X_n] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Besides,  $\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ . (**Standard error** is the standard deviation of an estimator)

- $\bar{X}$  is a linear combination of  $X_1, \dots, X_n$ .
- $\mathbb{E}[\bar{X}] = \mu$  and  $\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$  are regardless of the distribution of  $X$ .

### 3.5 Measuring quality of an estimator

- Let  $\psi(\theta)$  be any real valued function of  $\theta$ , suppose  $T$  is an estimator of  $\psi(\theta)$ . The most commonly used measurement of **accuracy** of an estimator is **mean squared error**,  $\text{MSE}_\theta(T) = \mathbb{E}_\theta[(T - \psi(\theta))^2]$ .
  - The smaller the value of  $\text{MSE}_\theta(T)$ , the more concentrated the sampling distribution of  $T$  is about the value  $\psi(\theta)$ .
  - Since the true value of  $\theta$  is unknown, often we evaluate the  $\text{MSE}_\theta(T)$  at  $\theta = \hat{\theta}$ .
- $\text{MSE}_\theta(T) = \text{Var}_\theta[T] + (\mathbb{E}_\theta[T] - \psi(\theta))^2$ .

*Proof.*

$$\begin{aligned}\text{MSE}_\theta(T) &= \mathbb{E}_\theta[(T - \psi(\theta))^2] = \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T] + \mathbb{E}_\theta[T] - \psi(\theta))^2] \\ &= \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2] + \mathbb{E}_\theta[(\mathbb{E}_\theta[T] - \psi(\theta))^2] + 2\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])(\mathbb{E}_\theta[T] - \psi(\theta))].\end{aligned}$$

We know

$$\begin{aligned}\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])(\mathbb{E}_\theta[T] - \psi(\theta))] &= \mathbb{E}_\theta[T - \mathbb{E}_\theta[T]](\mathbb{E}_\theta[T] - \psi(\theta)) \\ &= (\mathbb{E}_\theta[T] - \mathbb{E}_\theta[T])(\mathbb{E}_\theta[T] - \psi(\theta)) = 0.\end{aligned}$$

Besides,  $\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2] = \text{Var}_\theta[T]$ , and thus  $\text{MSE}_\theta(T) = \text{Var}_\theta[T] + (\mathbb{E}_\theta[T] - \psi(\theta))^2$ .  $\square$

### 3.6 Unbiasedness

- The bias of an estimator  $T$  of  $\psi(\theta)$  is given by  $\mathbb{E}_\theta[T] - \psi(\theta)$ .
- When the bias of an estimator is zero, it is called unbiased, i.e.,  $T$  is unbiased estimator of  $\psi(\theta)$  when  $\mathbb{E}_\theta[T] = \psi(\theta)$ . In other words,  $T$  is unbiased if  $\psi(\theta)$  is the mean of the sampling distribution of  $T$ .
- $\text{MSE}_\theta(T) = \text{Var}_\theta[T] + (\text{Bias}(T))^2$ .
  - For unbiased estimators,  $\text{MSE}_\theta(T) = \text{Var}_\theta[T]$ .
  - If all the other properties are similar, then an unbiased estimator is preferred over a biased estimator.

## 4 Sampling Distribution of $S^2$

### 4.1 Sample variance ( $S^2$ )

- Population variance:  $\sigma^2 = \mathbb{E}[(X - \mu)^2]$ , where  $\mu = \mathbb{E}[X]$ . If we have equally likely  $N$  data points in population,  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$ .
- $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$ .

*Proof.* We have

$$\begin{aligned} \sum_i (X_i - \mu)^2 &= \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_i (X_i - \bar{X})^2 + \sum_i (\bar{X} - \mu)^2 + 2 \sum_i (X_i - \bar{X})(\bar{X} - \mu) \\ &= \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_i (X_i - \bar{X}). \end{aligned}$$

We know

$$\sum_i (X_i - \bar{X}) = \sum_i X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0.$$

Therefore,

$$\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2. \quad \square$$

□

- Biased and unbiased estimator of  $\sigma^2$ : We have  $\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ , then we take expectation on both sides and have

$$\begin{aligned} \mathbb{E} \left[ \sum_i (X_i - \bar{X})^2 \right] &= \mathbb{E} \left[ \sum_i (X_i - \mu)^2 \right] - \mathbb{E} [n(\bar{X} - \mu)^2] \\ &= \sum_i \mathbb{E} [(X_i - \mu)^2] - n\mathbb{E} [(\bar{X} - \mu)^2] \\ &= \sum_i \text{Var}[X_i] - n\text{Var}[\bar{X}] \\ &= \sum_i \sigma^2 - n \frac{\sigma^2}{n} = (n - 1)\sigma^2. \end{aligned}$$



Therefore,  $\mathbb{E} \left[ \frac{1}{n} \sum_i (X_i - \bar{X})^2 \right] = \frac{n-1}{n} \sigma^2$ ,  $\mathbb{E} \left[ \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \right] = \sigma^2$ , i.e.,  $\frac{1}{n} \sum_i (X_i - \bar{X})^2$  is a biased estimator of  $\sigma^2$ ,  $\frac{1}{n-1} \sum_i (X_i - \bar{X})^2$  is an unbiased estimator of  $\sigma^2$ .

◦ For Normal distribution, both method of moments and MLE gives  $\frac{1}{n} \sum_i (X_i - \bar{X})^2$  as an estimator of  $\sigma^2$ .

◦  $\frac{n-1}{n} \rightarrow 1$  as  $n \rightarrow \infty$ , i.e., for large  $n$  both estimators will produce similar estimate.

◦ We choose  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

## 4.2 Sampling distribution of $S^2$ under Normal distribution

- Though the expression of  $S^2$  contains  $\bar{X}$ , they are independent. Besides, we can see a relation between  $S^2$  and  $\chi^2$  distribution.

**Theorem 4.1.** Suppose  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ ,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , then  $\bar{X} \perp S^2$ , and  $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2$ .

*Proof.*

**Lemma 1.** Suppose  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ ,  $U$  and  $V$  are two different linear combinations of the  $X_i$ ,  $\text{cov}[U, V] = 0$  iff  $U \perp V$ .

We know  $\bar{X} = \frac{1}{n} X_1 + \dots + \frac{1}{n} X_n$ ,  $X_1 - \bar{X} = (1 - \frac{1}{n}) X_1 - \frac{1}{n} X_2 - \dots - \frac{1}{n} X_n$ .

Besides,  $\text{cov}[\bar{X}, X_1 - \bar{X}] = \text{cov}[\bar{X}, X_1] - \text{cov}[\bar{X}, \bar{X}] = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$ . Similarly,  $\text{cov}[\bar{X}, X_i - \bar{X}] = 0, \forall i = 1, \dots, n$ .

By the Lemma, we know  $\bar{X} \perp X_i - \bar{X}$ , and thus

$$\bar{X} \perp \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = S^2.$$

Since  $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$ , then

$$\frac{\sum_i (X_i - \mu)^2}{\sigma^2} = \frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2},$$

i.e.,

$$\sum_i \left( \frac{X_i - \mu}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2.$$

Since  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , then  $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ , and  $\sum_i \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_{(n)}^2$ .

Since  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ , and  $\sum_i \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi_{(1)}^2$ . Besides, we have  $S^2 \perp \bar{X}$ , and therefore, we have

$$(1 - 2t)^{-\frac{n}{2}} = M_{\frac{(n-1)S^2}{\sigma^2}}(t) \cdot (1 - 2t)^{-\frac{1}{2}},$$

i.e,  $M_{\frac{(n-1)S^2}{\sigma^2}}(t) = (1 - 2t)^{-\frac{n-1}{2}}$ , and thus  $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2$ .  $\square$

- The mean of a  $\chi^2$  distribution is its df, then by theorem, we have  $\mathbb{E}\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$ , i.e.,  $\mathbb{E}[S^2] = \sigma^2$ . Hence,  $S^2$  is an unbiased estimator for  $\sigma^2$  under Normal distribution.
- An example of  $\text{cov} = 0 \not\Rightarrow$  independence.

**Example 4.1.**  $X \sim \mathcal{N}(0, 1)$ ,  $Y = X^2$ ,  $X$  and  $Y$  are dependent. However,

$$\text{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = \mathbb{E}[X^3] = 0.$$

### 4.3 $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

- We know  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ ,  $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2$ , and  $\bar{X} \perp S^2$ , then

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}.$$

#### 4.4 $\chi_{(m)}^2$

- $\chi_{(m)}^2 \sim \text{Gamma}\left(\frac{m}{2}, \frac{1}{2}\right)$ .  
 ◦ Gamma distribution:  $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ .
- $\frac{\chi_{(m)}^2}{m} = \frac{1}{m}(Z_1^2 + \dots + Z_m^2) = \frac{1}{m} \sum_{i=1}^m Z_i^2$ , where  $Z_i \sim \mathcal{N}(0, 1)$ . By LLN,

$$\frac{1}{m} \sum_{i=1}^m Z_i^2 \xrightarrow{P} \mathbb{E}[Z_i^2] = 1,$$

as  $m \rightarrow \infty$ .

- $t_{(m)} \xrightarrow{D} Z$ , as  $m \rightarrow \infty$ .

## 5 Properties of an Estimator: Consistency, Efficiency and Sufficiency

## 6 Interval Estimation

### 6.1 Confidence interval

- An interval  $C(X_1, \dots, X_n) = (l(X_1, \dots, X_n), u(X_1, \dots, X_n))$  is a  **$\gamma$ -confidence interval** for  $\psi(\theta)$  if  $P_\theta[\psi(\theta) \in C(X_1, \dots, X_n)] \geq \gamma, \forall \theta \in \Omega$ .  $\gamma$  represents the confidence level of the interval.

◦ In naive words: We want two numbers which will have at least  $\gamma$  chance of containing the true parameter.

### 6.2 CI for parameters of Normal distribution

#### 6.2.1 CI for $\mu$ with $\sigma^2$ known

- We know  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ , we can write

$$P\left[k_1 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq k_2\right] \geq \gamma \Rightarrow P\left[\bar{X} - k_2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - k_1 \frac{\sigma}{\sqrt{n}}\right] \geq \gamma.$$

- $k_1$  and  $k_2$  are quantiles of  $\mathcal{N}(0, 1)$  s.t.  $P[k_1 \leq Z \leq k_2] \geq \gamma$ .
- The sampling distribution is unimodal and symmetric around the mode, the middle  $\gamma$  part gives the shortest interval and thus  $z_{\frac{1-\gamma}{2}}$  and  $z_{\frac{1+\gamma}{2}}$  are preferred as the value of  $k_1$  and  $k_2$ . For example, if  $\gamma = 0.95$ ,  $k_1 = z_{0.025} = -1.96$ ,  $k_2 = z_{0.975} = 1.96$ .
- For  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known, the  $\gamma$ -CI of  $\mu$  is

$$\left[\bar{X} - z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

#### 6.2.2 CI for $\mu$ with $\sigma^2$ unknown

- When  $\sigma^2$  is unknown, we use  $S^2$  as an estimator of  $\sigma^2$  and we have  $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$ .
- For  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  unknown, the  $\gamma$ -CI of  $\mu$  is

$$\left[\bar{X} - t_{\frac{1+\gamma}{2}(n-1)} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{1+\gamma}{2}(n-1)} \frac{S}{\sqrt{n}}\right],$$

where  $t_{\frac{1+\gamma}{2}(n-1)}$  is the  $\frac{1+\gamma}{2}$  quantile of a  $t_{(n-1)}$  distribution.

### 6.2.3 CI for $\sigma^2$

- We know  $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ , we can write

$$P \left[ \chi^2_{\frac{1-\gamma}{2}(n-1)} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\frac{1+\gamma}{2}(n-1)} \right] \geq \gamma \Rightarrow P \left[ \frac{(n-1)S^2}{\chi^2_{\frac{1+\gamma}{2}(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{1-\gamma}{2}(n-1)}} \right] \geq \gamma.$$

- For  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , the  $\gamma$ -CI of  $\sigma^2$  is  $\left[ \frac{(n-1)S^2}{\chi^2_{\frac{1+\gamma}{2}(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{1-\gamma}{2}(n-1)}} \right]$ .
- Remark:
  - $\chi^2$  is not a symmetric distribution (at least for lower df).
  - The shape of  $\chi^2$  depends on its df.
  - Using  $\chi^2_{\frac{1+\gamma}{2}(n-1)}$  and  $\chi^2_{\frac{1-\gamma}{2}(n-1)}$  as two ends may not result in the shortest length.

## 6.3 CI for mean of a non-Normal distribution using CLT

- The  $\gamma$ -CI of  $\mu$  is  $\left[ \bar{X} - z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}} \right]$ ,  $\sigma^2$  may be unknown.
  - If  $\sigma^2$  is unknown, we can use MLE to calculate  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ .

**Example 6.1.** CI for  $\lambda$  when data follows  $\text{Poisson}(\lambda)$ .

*Solution.* By CLT,  $\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{D} \mathcal{N}(0, 1)$ , where  $\text{SE}(\bar{X}) = \sqrt{\frac{\lambda}{n}}$ . We know  $\bar{X}$  is the MLE of  $\lambda$ , then the estimated  $\text{SE} = \sqrt{\frac{\bar{X}}{n}}$ . Thus, the  $\gamma$ -CI for  $\lambda$  is  $\left[ \bar{X} - z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{X}}{n}}, \bar{X} + z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{X}}{n}} \right]$ .

## 6.4 Interpreting CI

- For  $z$  and  $t$  interval, the sample mean  $\bar{X}$  is the midpoint of the lower and upper bound.

- Width of the interval = Upper bound–Lower bound. Half of the width is known as the *margin of error* (ME). CI:  $[\bar{X} \pm \text{ME}]$ .
  - $\gamma \uparrow \Rightarrow$  Width of the interval  $\uparrow$ .
  - $\sigma$  or  $s \uparrow \Rightarrow$  Width of the interval  $\uparrow$ .
  - $n \uparrow \Rightarrow$  Width of the interval  $\downarrow$ .
- Interpretation: If we keep taking samples (infinite times) and keep constructing  $\gamma$ -CIs, in  $100\gamma\%$  of the cases, our CIs will capture the true value of the parameter.

## 7 Test of Hypothesis

### 7.1 Types of hypothesis

- **Null hypothesis**/ $H_0$  : The hypothesis that we want to test.
- **Alternative hypothesis**/ $H_A/H_1$  : The alternative values of the parameter of interest.
  - Often this is what we are trying to prove as a researcher.
- **Simple hypothesis**: When a hypothesis involves only a single value from the parameter space.
- **Composite hypothesis**: When a hypothesis involves more than one values from the parameter space.
- In practice, often we test **simple null** hypothesis against **composite alternative** hypothesis.

### 7.2 Two approaches of hypothesis testing

#### 7.2.1 Critical region approach

- Due to uncertainty, often we reject  $H_0$  even though it could be true. We assign a preferably small predefined probability of making this mistake and call it **level of significance**, denoted by  $\alpha$ .
- **Test statistic**,  $T(X)$ , is a quantity that simultaneously serves few purposes:
  - It summarizes the sample data through an estimator.
  - When  $H_0$  is true, it has a known distribution.
  - Under that distribution, it is possible to find some areas that has probability  $\alpha$ .
- **Critical region**,  $R_\alpha(T)$ , is a region of the distribution of the test statistic s.t. we will reject  $H_0$  if  $T(X) \in R_\alpha(T)$ . We need  $P[T(X) \in R_\alpha(T) | H_0 \text{ is true}] = \alpha$ .



- Testing  $H_0 : \mu = \mu_0$  when  $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known:
  - $H_0 : \mu = \mu_0$ .
  - $T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ .
  - If  $H_0$  is true, then  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ .
  - Rejection region:  $(-\infty, z_{\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$ .
  - We reject  $H_0$  if  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}$  or  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\frac{\alpha}{2}}$ .
  - Intuition: We reject the null hypothesis when the test statistic falls in the lower probability area of the distribution under the null. In naive words: If  $\mu_0$  is the true mean, then  $\bar{X}$  should not be too far from  $\mu_0$ .
  - Note: We never say we accept  $H_0$ . We failed to prove that  $H_0$  is wrong  $\nRightarrow H_0$  is right.
- Testing  $H_0 : \mu = \mu_0$  when  $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  unknown:
  - $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$ .
  - Rejection region:  $(-\infty, t_{\frac{\alpha}{2}(n-1)}) \cup (t_{1-\frac{\alpha}{2}(n-1)}, \infty)$ .
- Testing  $H_0 : \sigma^2 = \sigma_0^2$  when  $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ :
  - $T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{(n-1)}^2$ .
  - $R_\alpha(T) = (-\infty, \chi_{\frac{\alpha}{2}(n-1)}^2) \cup (\chi_{1-\frac{\alpha}{2}(n-1)}^2, \infty)$ .

### 7.2.2 $p$ -value approach

- $p$ -value: It is the smallest level of significance at which  $H_0$  would be rejected based on the observed data. Also, it is the probability of observing the result as or more extreme than that actually observed if  $H_0$  is true. In naive words:  $p$ -value suggests how surprising the observed sample is if we assume  $H_0$  to be true.
  - Conventionally, we compare  $p$ -value to 0.01, 0.05 or 0.1.
  - If  $p$ -value is less than a predefined cut-off, we reject  $H_0$ .

- For  $z$ -test,  $p$ -value =  $2 \left[ 1 - \Phi \left( \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \right) \right]$ .
- For  $t$ -test,  $p$ -value =  $2 \left[ 1 - G \left( \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \right) \right]$ , where  $G$  is the CDF of a  $t_{(n-1)}$  distribution.

### 7.3 Type-1, 2 error and power of a test

- Definition
  - $P[\text{Type} - 1 \text{ error}] = \alpha = P[\text{Reject } H_0 | H_0 \text{ is true}]$ .
  - $P[\text{Type} - 2 \text{ error}] = \beta = P[\text{Fail to reject } H_0 | H_0 \text{ is false}]$ .
  - Power of a test =  $1 - \beta = P[\text{Reject } H_0 | H_0 \text{ is false}]$ .
- Graph analysis: Suppose we are testing two simple hypotheses,  $H_0 : \mu = 1, H_1 : \mu = 4$ , and there are no other options. The area shaded in red is type-1 error and in cyan is type-2 error.

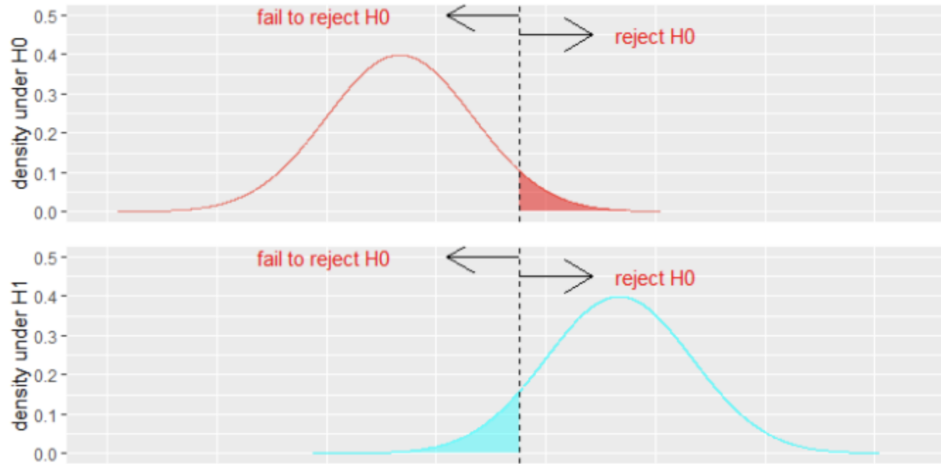


Figure 7.1:  $H_0 : \mu = 1, H_1 : \mu = 4$ .

**Example 7.1.** Suppose we have  $\mathcal{N}(\mu, \sigma^2)$  populations with unknown  $\mu$  and  $\sigma = 3$ . We want to test  $H_0 : \mu = 1, H_1 : \mu = 4$  at  $\alpha = 0.05, n = 9$ . Calculate  $\beta$  and  $1 - \beta$ .

*Solution.* We have  $\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = 1$ .

Therefore, under  $H_0$ ,  $\bar{X} \sim \mathcal{N}(1, 1)$  and under  $H_1$ ,  $\bar{X} \sim \mathcal{N}(4, 1)$ . Hence,  $R_\alpha = \frac{\bar{X}-1}{1} > z_{0.95} \Rightarrow \bar{X} > 2.645$ .

Therefore,

$$1 - \beta = P[\bar{X} > 2.645 | H_1] = P\left[\frac{\bar{X} - 4}{1} > \frac{2.645 - 1}{1}\right] = 0.912,$$

and  $\beta = 1 - 0.912 = 0.088$ .

## 7.4 Test of hypothesis using CI

- Let  $\alpha = 1 - \gamma$ . Constructing a  $\gamma$  level CI for  $\mu$  and checking whether  $\mu_0$  is inside or not is equivalent of testing the hypothesis of  $\mu = \mu_0$  at  $(1 - \gamma)$  level of significant.

## 8 Likelihood Ratio Test and Comparing Two Populations

### 8.1 Likelihood ratio test (LRT)

- General definition: Suppose we are testing  $H_0 : \theta \in \Omega_0, H_1 : \theta \in \Omega_1$ . Let  $L(\theta)$  represents the likelihood function. The generalized likelihood ratio is defined as  $\Lambda^* = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega_1} L(\theta)}$ . A small value of  $\Lambda^*$  provides evidence against  $H_0$ .
- Special case:  $\Lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} = \frac{\max_{\theta \in \Omega_0} L(\theta)}{L(\hat{\theta})}$ , where  $\hat{\theta}$  is MLE of  $\theta$ .
  - If  $\hat{\theta} \in \Omega_0$ , then  $\Lambda = 1 \Rightarrow$  we will not reject  $H_0$ .
  - If  $\hat{\theta} \notin \Omega_0$ , we look for the most likely  $\theta$  value in  $\Omega_0$  and check if it does a good enough job as it is done by the MLE.
  - $\Lambda$  value closer to 0 will provide evidence against  $H_0$ .

**Theorem 8.1.** Let  $p = \dim \Omega$  be the number of free parameters in the whole parameter space,  $d = \dim \Omega_0$  be the number of free parameters under the null, then we have  $-2 \ln \Lambda \xrightarrow{P} \chi^2_{(p-d)}$ , when  $H_0$  is true.

**Example 8.1.**  $(X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$ . Test  $H_0 : \mu = \mu_0$  at level of significance  $\alpha$ .

*Solution.* We have  $L(\mu) = (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_0^2} \sum (X_i - \mu)^2 \right]$ .

Under  $H_0$ ,  $L(\mu_0) = (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_0^2} \sum (X_i - \mu_0)^2 \right]$ .

We know  $L(\mu)$  is maximized at  $\bar{X}$  and thus

$$L(\hat{\mu}) = (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_0^2} \sum (X_i - \bar{X})^2 \right].$$

Therefore,

$$\begin{aligned} \Lambda &= \frac{L(\mu_0)}{L(\hat{\mu})} = \exp \left[ -\frac{1}{2\sigma_0^2} \left( \sum (X_i - \mu_0)^2 - \sum (X_i - \bar{X})^2 \right) \right] \\ &= \exp \left[ -\frac{1}{2\sigma_0^2} n(\bar{X} - \mu_0)^2 \right]. \end{aligned}$$

Besides,  $p = 1, d = 0$  and thus

$$-2 \ln \Lambda = \frac{1}{\sigma_0^2} n (\bar{X} - \mu_0)^2 = \left( \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right)^2 \sim \chi_{(1)}^2.$$

We reject  $H_0$  if  $-2 \ln \Lambda > \chi_{1-\alpha(1)}^2$ .

- LRT for non-Normal distribution: LRT allows us to test hypothesis for non-Normal distributions since all we need is the likelihood function evaluated at  $\theta_0$  and  $\hat{\theta}$ .

**Example 8.2.** Suppose  $X_i \sim \text{Exp}(\theta), \mathbb{E}[X] = \theta$ . We test  $H_0 : \theta = 60, H_1 : \theta \neq 60$ . Besides,  $n = 100, \bar{x} = 75$ .

*Solution.* (Method 1)  $L(\theta) = \frac{1}{\theta^n} \exp \left[ -\frac{1}{\theta} \sum_{i=1}^n X_i \right]$  and the MLE is  $\bar{X}$ .

Therefore,  $\Lambda = \left( \frac{\bar{X}}{\theta_0} \right)^n \exp \left[ n \left( 1 - \frac{\bar{X}}{\theta_0} \right) \right]$  and thus

$$-2 \ln \Lambda = -2n \left( \ln \bar{X} - \ln \theta_0 + 1 - \frac{\bar{X}}{\theta_0} \right) \sim \chi_{(1)}^2.$$

Since  $\theta_0 = 60, n = 100, \bar{x} = 75$ , then  $-2 \ln \Lambda = 5.37 > \chi_{0.95(1)}^2 = 3.84$ . Thus we reject  $H_0$  at  $\alpha = 0.05$ .

(Method 2) If  $H_0$  is true, then  $-2 \ln \Lambda \sim \chi_{(1)}^2$  and  $p\text{-value} = P(\chi_{(1)}^2 > 5.37) = 0.02$ .

## 8.2 Constructing CI using LRT

- Under  $H_0, -2 \ln \Lambda \xrightarrow{D} \chi_{(p-d)}^2$ , we reject  $H_0$  if  $-2 \ln \Lambda > \chi_{1-\alpha(p-d)}^2$ . Conversely, we will fail to reject if  $-2 \ln \Lambda < \chi_{1-\alpha(p-d)}^2$ . Thus,  $(1 - \alpha)$  level CI for  $\theta$  is the interval of  $\theta$  values for which  $-2 \ln \Lambda < \chi_{1-\alpha(p-d)}^2$ , i.e.,  $L(\theta) > L(\hat{\theta}) \exp \left[ -\frac{\chi_{1-\alpha(p-d)}^2}{2} \right]$ .

## 8.3 Comparing two independent Normal population

### 8.3.1 Equality of two variances

- Suppose we have two independent Normal samples  $X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$  and  $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ . We want to test  $H_0 : \sigma_X^2 = \sigma_Y^2$ , and  $H_1 : \sigma_X^2 \neq \sigma_Y^2$ .
- We have  $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{(n-1)}^2$ ,  $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{(m-1)}^2$  and thus

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{(n-1, m-1)}.$$

Under  $H_0$ , we have  $\frac{S_X^2}{S_Y^2} \sim F_{(n-1, m-1)}$ .

- The rejection region is  $\left(-\infty, F_{\frac{\alpha}{2}(n-1, m-1)}\right) \cup \left(F_{1-\frac{\alpha}{2}(n-1, m-1)}, \infty\right)$ .

### 8.3.2 Equality of two means with variances known

- We want to test  $H_0 : \mu_X = \mu_Y$ , which is same to test  $H_0 : \mu_X - \mu_Y = 0$ .
- We have  $\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n})$ ,  $\bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y^2}{m})$  and thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0, 1).$$

Under  $H_0$ , we have

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0, 1).$$

- The  $(1 - \alpha)$  level CI is  $\left[(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right]$  and check if 0 is inside or not. Or, the rejection region is  $\left(-\infty, z_{\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right)$ . Or, calculate the  $p$ -value.
- If  $\sigma_X = \sigma_Y = \sigma$ , then under  $H_0$ , we have  $\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0, 1)$ .

### 8.3.3 Equality of two means with variances unknown

- Suppose  $\sigma_X = \sigma_Y = \sigma$ .
- We have  $\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0, 1)$ , and

$$\begin{aligned} \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} &= \frac{1}{\sigma^2}[(n-1)S_X^2 + (m-1)S_Y^2] \\ &\sim \chi_{(n-1)}^2 + \chi_{(m-1)}^2 = \chi_{(n+m-2)}^2. \end{aligned}$$

Therefore,

$$\frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\frac{1}{\sigma^2}[(n-1)S_X^2 + (m-1)S_Y^2]/(n+m-2)}} \sim t_{(n+m-2)},$$

i.e.,

$$\frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(n+m-2)},$$

where  $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$  is called the *pooled sample variance*.

## 8.4 Comparing two population means (paired data)

- In many practical setting, the samples are paired and thus the observations are not independent.
- We want to test  $H_0 : \mu_X - \mu_Y = 0$ ,  $H_1 : \mu_X - \mu_Y \neq 0$ .
  - If we use  $\bar{X} - \bar{Y}$ ,  $\text{Var}[\bar{X} - \bar{Y}]$  will contain a covariance term.
  - To simplify, define  $D = X - Y \Rightarrow \mu_D = \mu_X - \mu_Y$ , and thus

$$\frac{\bar{D}}{S_D/\sqrt{n}} \sim t_{(n-1)}.$$

## 8.5 Comparing two populations using LRT

- Suppose we have two independent Normal samples:  $X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$  and  $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ , where  $\sigma_X^2$  and  $\sigma_Y^2$  are known. We want to test  $H_0 : \mu_X = \mu_Y$  by LRT.

◦ We have two unknown parameters  $\mu_X, \mu_Y$ . Under  $H_0$ ,  $\mu_X = \mu_Y = \mu$ , then we have one unknown parameter.

◦ We have

$$L(\mu_X, \mu_Y) = (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_X^2} \sum_{i=1}^n (X_i - \mu_X)^2 \right] (2\pi\sigma_Y^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_Y^2} \sum_{i=1}^n (Y_i - \mu_Y)^2 \right],$$

and  $\hat{\mu}_X = \bar{X}, \hat{\mu}_Y = \bar{Y}$ .

◦ Under  $H_0$ , we have

$$L(\mu) = (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_X^2} \sum_{i=1}^n (X_i - \mu)^2 \right] (2\pi\sigma_Y^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_Y^2} \sum_{i=1}^n (Y_i - \mu)^2 \right],$$

and to find the MLE of  $\mu$ , we have

$$l(\mu) = C - \frac{1}{2\sigma_X^2} \sum (X_i - \mu)^2 - \frac{1}{2\sigma_Y^2} \sum (Y_j - \mu)^2.$$

Hence,

$$\partial_\mu l = \frac{1}{\sigma_X^2} \sum (X_i - \mu) + \frac{1}{\sigma_Y^2} \sum (Y_j - \mu) = \frac{1}{\sigma_X^2} (n\bar{X} - n\mu) + \frac{1}{\sigma_Y^2} (m\bar{Y} - m\mu).$$

Let  $\partial_\mu l = 0$ , we have

$$\hat{\mu} = \frac{\frac{1}{\sigma_X^2/n}}{\frac{1}{\sigma_X^2/n} + \frac{1}{\sigma_Y^2/m}} \bar{X} + \frac{\frac{1}{\sigma_Y^2/m}}{\frac{1}{\sigma_X^2/n} + \frac{1}{\sigma_Y^2/m}} \bar{Y}.$$

◦ Hence,  $-2 \ln \Lambda = -2 \ln \frac{L(\hat{\mu})}{L(\hat{\mu}_X, \hat{\mu}_Y)}$  and under  $H_0$ ,  $-2 \ln \Lambda \sim \chi_{(1)}^2$ .

## 8.6 Numerical example

**Example 8.3.**  $(4, 10, 10, 4, 6, 8, 8, 3, 4, 4) \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ . Test  $H_0 : \lambda = 5$ .

*Solution.* (Method 1)  $L(\lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$ . Since  $n = 10, \lambda_0 = 5, \hat{\lambda} = \bar{x} = 6.1$ , then we have

$$\Lambda = \frac{e^{-50} 5^{61}}{e^{-61} (6.1)^{61}} = 0.3231, -2 \ln \Lambda = 2.2598.$$



Since  $\chi^2_{0.95(1)} = 3.841459$ ,  $-2 \ln \Lambda < \chi^2_{0.95(1)}$ , then we fail to reject  $H_0$ .

(Method 2) If  $H_0$  is true, then  $-2 \ln \Lambda \sim \chi^2_{(1)}$ . Thus,  $p\text{-value} = P[\chi^2_{(1)} > 2.2598] = 0.13 > 0.05$ .

**Example 8.4.** (Rice, pp.425, B)  $\bar{x}_A = 80.02$ ,  $\bar{x}_B = 79.98$ ,  $s_{x_A} = 0.024$ ,  $s_{x_B} = 0.031$ , and  $\sigma_A, \sigma_B$  are unknown.

*Solution.* We have  $s_p^2 = \frac{12(0.024)^2 + 7(0.031)^2}{19}$ ,  $s_p \sqrt{\frac{1}{n} + \frac{1}{m}} = 0.012$ .

The test statistic is  $T = 3.3333$ ,  $t_{0.975(19)} = 2.093$ . Since  $T > t_{0.975(19)}$ , we reject  $H_0$ . The 95% CI for  $\mu_{x_A} - \mu_{x_B}$  is  $\left[ (\bar{x}_A - \bar{x}_B) \pm t_{0.975(19)} s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right] = [0.015, 0.065]$ .

**Example 8.5.** (Week 8 slide, pp. 32) Let  $X$  and  $Y$  represent the before and after measurements of 10 participants. Check whether the drink changes the blood sugar level or not.

*Solution.* We have  $\bar{d} = 4.47$ ,  $s_d = 3.545106$ .

The test statistic is  $T = \frac{\bar{d}}{s_d/\sqrt{n}} = 3.987294$ ,  $t_{0.975(9)} = 2.262$ . Since  $T > t_{0.975(9)}$ , we reject  $H_0$ . Besides, the rejection region is  $(-\infty, -2.262) \cup (2.262, \infty)$ .

## 9 Model Checking

### 9.1 $\chi^2$ goodness of fit test

- The test is used to assess whether or not a **categorical random variable**  $W$ , which takes finite values  $\{1, 2, \dots, k\}$ , has a specified probability measure  $P$ .
  - When we have discrete r.v. which takes infinitely many values, we partition the possible values into  $k$  categories.
  - When we have a continuous r.v., we partition the real line into  $k$  sub-intervals.
  - Naturally, the counts of these  $k$  categories form a **multinomial distribution**.

**Theorem 9.1.** Let  $X_1, \dots, X_k$  be the observed counts of category  $1, 2, \dots, k$  respectively. We can write  $(X_1, \dots, X_k) \sim \text{Multinomial}(n, p_1, \dots, p_k)$ , where  $p_1, \dots, p_k$  are **known**, and we have

$$\mathbb{E}[X_i] = np_i, \text{Var}[X_i] = np_i(1 - p_i).$$

The test statistic  $T$  is

$$X^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \xrightarrow{D} \chi_{(k-1)}^2.$$

Or we can say

$$X^2 = \sum_{i=1}^k \frac{(\text{Observed count of } i - \text{Expected count of } i)^2}{\text{Expected count of } i} \xrightarrow{D} \chi_{(k-1)}^2.$$

*Proof.* (For the simple case, i.e.,  $k = 2$ ). We have

$$\begin{aligned} X^2 &= \sum_{i=1}^2 \frac{(X_i - np_i)^2}{np_i} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1 - p_1))^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{n} \left( \frac{1}{p_1} + \frac{1}{p_2} \right) = \left( \frac{X_1 - np_1}{\sqrt{np_1 p_2}} \right)^2 \xrightarrow{D} \chi_{(1)}^2. \end{aligned}$$

□

- It is recommended to ensure that  $\mathbb{E}[X_i] = np_i \geq 1, \forall i$ .

**Example 9.1.** Suppose we have 10000 random numbers generated from a Uniform[0, 1] distribution. After dividing them into 10 equal length bins, we test if these numbers look uniform or not.

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	993	1044	1061	1021	1017	973	975	965	996	955

*Solution.* If the numbers are really from a Uniform[0, 1] distribution then expected counts for each cell is  $10000 \cdot \frac{1}{10} = 1000$ , so we have

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	993	1044	1061	1021	1017	973	975	965	996	955
$\hat{x}_i$	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

The test statistic is  $X^2 = \frac{(993-1000)^2}{1000} + \dots + \frac{(955-1000)^2}{1000} = 11.056$ . The  $p$ -value is 0.27189, and thus we fail to reject the statement that these number are from a Uniform[0, 1] distribution. In naive words, they look uniform, and the code for  $p$ -value is:

```
1 1 - pchisq(11.056, 9)
```

Or:

```
1 x = c(993, 1044, 1061, 1021, 1017, 973, 975, 965, 996, 955)
2 chisq.test(x)
```

**Remark.** Since we divided the range [0, 1] into 10 bins and we know it is uniform,  $p_k$ 's are all 0.1, which are constants and do not need to be estimated using any of the sample observations.

**Theorem 9.2.** If  $p_1, \dots, p_k$  are **unknown**, then we need to estimate them. In this case  $(X_1, \dots, X_k) \sim \text{Multinomial}(n, p_1(\theta), \dots, p_k(\theta))$ . After estimating  $\theta$  by  $\hat{\theta}$ , the test statistic is

$$X^2 = \sum_{i=1}^k \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow{D} \chi_{(k-1-\dim \Omega)}^2,$$

where  $\dim \Omega$  represents the number of parameters needed to be estimated based on the data in order to calculate the  $p_i$ 's.

**Example 9.2.** Suppose life-lengths of light bulbs ( $Y_i$ ) follows an Exponential( $\beta$ ), where  $\beta$  is unknown. We have the partitions as

$$(0, 1], (1, 2], (2, 3], (3, \infty).$$

Based on the sample of size  $n = 30$ , the observed counts are 5, 16, 8, 1. We test  $H_0$  : The true model is Exponential( $\beta$ ).

*Solution.* First, we find the MLE for  $\beta$ . If the life-lengths of the 30 bulbs are available, then

$$L(\beta) = \beta^{30} \exp \left[ -\beta \sum y_i \right] \Rightarrow \hat{\beta} = \frac{1}{\bar{y}}.$$

If all we have is the counts of  $Y_i$ 's that fall into those four partitions, we can define

$$L(\beta) = (1 - e^{-\beta})^2 (e^{-\beta} - e^{-2\beta})^{16} (e^{-2\beta} - e^{-3\beta})^8 (e^{-3\beta})^1,$$

where  $(1 - e^{-\beta}) = P(Y_i \in (0, 1])$ , similarly the other terms. For instance,

$$p_2 = \int_1^2 \beta e^{-\beta x} dx = e^{-\beta} - e^{-2\beta}.$$

Thus, we have  $\hat{\beta} = 0.603535$ , and

$$p_1 = 0.453125,$$

$$p_2 = 0.247803,$$

$$p_3 = 0.135517,$$

$$p_4 = 0.163555.$$

The expected counts are 13.59375, 7.43409, 4.06551, 4.90665, respectively.

Hence, the test statistic is  $X^2 = \frac{(5-13.59375)^2}{13.59375} + \dots = 22.22$ . The  $p$ -value is 0.000015, and thus we reject  $H_0$ , i.e., we have strong evidence that Exponential( $\beta$ ) is not the true model for these data and the code for  $p$ -value is:

1 `1 - pchisq(22.22, 2)`

**Remark.** Since we estimate  $\beta$  using given data, we loose 1 extra degrees of freedom, and thus it is  $\chi^2_{(2)}$ .

## 9.2 Discrepancy statistic

- Suppose  $(X_1, \dots, X_n)$  is believed to be from  $f_\theta$  with  $\theta \in \Omega$ . **Discrepancy statistic**,  $D(X)$  is a function that takes the samples observations and maps it to  $\mathbb{R}$ . It measures the deviation from the model under consideration. A large value of  $D(X)$  implies a deviation has occurred.
  - In test of hypothesis sense, we assess whether  $D(X)$  lies in the region of low probability of its distribution when the model is correct.
  - Restriction: When the model is correct,  $D$  must have a single distribution, i.e., the distribution of  $D$  cannot depend on  $\theta$ .
  - A statistic  $D$  whose distribution under the model does not depend upon  $\theta$  is called **ancillary**, i.e., if  $(X_1, \dots, X_n) \sim f_\theta$ , then  $D(X)$  has the same distribution for every  $\theta \in \Omega$ .
- \* Being ancillary does not mean  $D$  can be used as a discrepancy statistic.
- \* If  $D$  is constant, then it is ancillary, but not useful for model checking.

**Example 9.3.** Suppose  $(X_1, \dots, X_n) \sim \mathcal{N}(\mu, \sigma_0^2)$ ,  $X_i$ 's are independent. Define  $R_i = X_i - \bar{X}$ . For instance,

$$X_1 - \bar{X} = X_1 - \frac{1}{n}(X_1 + \dots + X_n) = (1 - \frac{1}{n})X_1 - \frac{1}{n}X_2 - \dots - \frac{1}{n}X_n.$$

Thus,

$$\mathbb{E}[X_1 - \bar{X}] = \mathbb{E}[X_1] - \mathbb{E}[\bar{X}] = \mu - \mu = 0,$$

and

$$\begin{aligned} \text{Var}[X_1 - \bar{X}] &= \text{cov}(X_1 - \bar{X}, X_1 - \bar{X}) \\ &= \text{cov}((1 - \frac{1}{n})X_1 - \frac{1}{n}X_2 - \dots - \frac{1}{n}X_n, (1 - \frac{1}{n})X_1 - \frac{1}{n}X_2 - \dots - \frac{1}{n}X_n) \\ &= (1 - \frac{1}{n})\sigma_0^2, \end{aligned}$$

Therefore,  $R_i \sim \mathcal{N}(0, (1 - \frac{1}{n})\sigma_0^2)$ . The discrepancy statistic

$$D(R) = \frac{1}{\sigma_0^2} \sum_{i=1}^n R_i^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{(n-1)}^2$$

If  $D(r)$  represent the observed value of  $D$  based on the current sample then, then we can calculate the  $p$ -value.

### 9.3 Residual and quantile/probability plots

- Residual plot: Since  $R_i \sim \mathcal{N}(0, (1 - \frac{1}{n})\sigma_0^2)$ , we can define **standardized residual**

$$r_i^* = \frac{x_i - \bar{x}}{\sqrt{(1 - \frac{1}{n})\sigma_0^2}}.$$

If the true model is  $\mathcal{N}(\mu, \sigma_0^2)$ , then our expectation is that  $r_i^*$ 's will behave like values from a  $\mathcal{N}(0, 1)$ .

- Plotting  $r_1^*, \dots, r_n^*$  against  $(1, \dots, n)$ .
- The points should be clustered around zero.
- The points should lie in  $(-3, 3)$ .
- They should look random (should not depict any pattern).

**Example 9.4.** Points in Figure 9.2 satisfies the conditions above. Some of points in Figure 9.3 are outside  $(-3, 3)$ , indicating longer tail. Most of points in Figure 9.4 are on positive side, indicating right skewed.

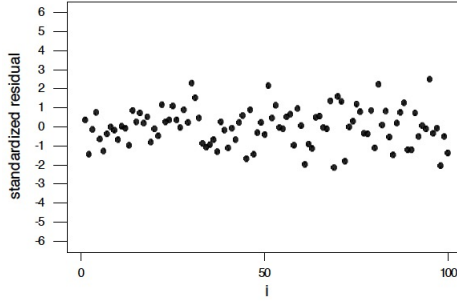


Figure 9.2: A plot of the standardized residuals for a sample of 100 from an  $\mathcal{N}(0, 1)$  distribution.

- Quantile/Probability plots: Suppose  $(X_i)$  is believed to be from  $\mathcal{N}(\mu, \sigma^2)$ . Let  $X_{(i)}$  represent the  $i$ -th order statistic. We have

$$\mathbb{E}[X_{(i)}] = \mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right),$$

where  $\Phi^{-1}$  is the inverse CDF of  $\mathcal{N}(0, 1)$ .

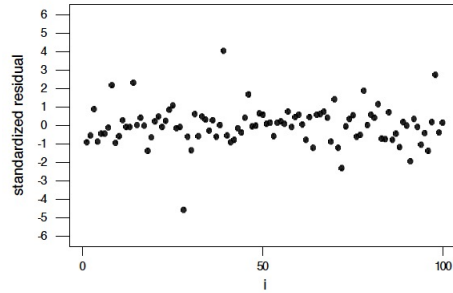


Figure 9.3: A plot of the standardized residuals for a sample of 100 from  $X = (\sqrt{3})^{-1}Z$ , where  $Z \sim t_{(3)}$ .

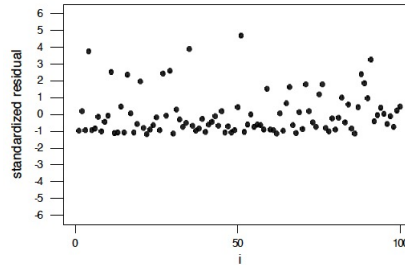


Figure 9.4: A plot of the standardized residuals for a sample of 100 from an  $\text{Exponential}(1)$  distribution.

Let  $x_j$  correspond to the order statistic  $x_{(i)}$ , then  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  is the **Normal score** of  $x_j$ . If we plot the points  $\left(x_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$ , they should lie approximately on a straight line with intercept  $\mu$  and slope  $\sigma$ .

**Example 9.5.** Suppose we want to assess whether or not the following data set can be considered a sample of sample of size  $n = 10$  from some Normal distribution:

2.00 0.28 0.47 3.33 1.66 8.17 1.18 4.15 6.43 1.77

The order statistics and associated Normal scores are

i	1	2	3	4	5
$x_{(i)}$	0.28	0.47	1.18	1.66	1.77
$\Phi^{-1}\left(\frac{i}{n+1}\right)$	-1.34	-0.91	-0.61	-0.35	-0.12
i	6	7	8	9	10
$x_{(i)}$	2.00	3.33	4.15	6.43	8.17
$\Phi^{-1}\left(\frac{i}{n+1}\right)$	0.11	0.34	0.60	0.90	1.33



## 10 $\chi^2$ Test of Independence and Homogeneity

### 10.1 Relationship among variables

- Variables  $X$  and  $Y$  are **related variables** if there is any change in the conditional distribution of  $Y$ , given  $X = x$ , as  $x$  changes.

**Example 10.1.** Assume  $Y \sim \mathcal{N}(10, 2)$  when  $X = 1$  and  $Y \sim \mathcal{N}(12, 2)$  when  $X = 0$ . Since the mean changes, any probability we calculate for  $Y$  will be different based on the value of  $X$ . Similarly, the variance of  $Y$  can be function of  $X$  as well. In this case, we can say  $X$  and  $Y$  are related, which means they are not independent.

- Often, we think of  $Y$  as the **dependent variable** (depending on  $X$ ) and  $X$  as the **independent variable** (free to vary). Also,  $Y$  is called the **response variable** and  $X$  is called the **predictor variable**.

### 10.2 Relationship of two categorical variables

- Assume we have two categorical variables and we want to check whether  $X$  and  $Y$  are related or not. Assume  $Y$  is the response and  $X$  is the predictor.  $X, Y$  has  $a, b$  number of categories respectively.

#### 10.2.1 $\chi^2$ test of independence ( $X$ and $Y$ are random)

- Notation:
  - Let  $i = 1, 2, \dots, a$  be the  $a$  categories of  $X$  and  $j = 1, 2, \dots, b$  be the  $b$  categories of  $Y$ .
  - Let  $f_{ij}$  be the number of samples corresponding to  $i$ th category of  $X$  and  $j$ th category of  $Y$ . We have  $\sum_{i=1}^a \sum_{j=1}^b f_{ij} = n$ .
  - Let  $F_{ij}$  be the population count of the  $(i, j)$ th cell.
  - Let  $\theta_{ij} = P(X = i, Y = j)$ , i.e., the proportion of elements in the population with  $X = i$  and  $Y = j$ . We can write

$$(F_{11}, F_{12}, \dots, F_{ab}) \sim \text{Multinomial}(n, \theta_{11}, \theta_{12}, \dots, \theta_{ab}).$$

Let  $\theta_{i.}$  be the marginal probability  $P(X = i)$  and  $\theta_{.j}$  be the marginal probability  $P(Y = j)$ .

- We want to test  $H_0$  : There is no relationship between  $X$  and  $Y \Rightarrow X \perp Y$ .

◦ If  $X \perp Y$ , then  $P(X = i, Y = j) = P(X = i)P(Y = j)$ , i.e., under  $H_0$ ,  $\theta_{ij} = \theta_{i.}\theta_{.j}$ , and thus

$$(F_{11}, F_{12}, \dots, F_{ab}) \sim \text{Multinomial}(n, \theta_{1.}\theta_{.1}, \theta_{1.}\theta_{.2}, \dots, \theta_{a.}\theta_{.b}).$$

- Test statistic and corresponding distribution:

◦ MLE of  $\theta_{i.}$  will be  $\hat{\theta}_{i.} = \sum_{j=1}^b \frac{f_{ij}}{n}$  and MLE of  $\theta_{.j}$  will be  $\hat{\theta}_{.j} = \sum_{i=1}^a \frac{f_{ij}}{n}$ .

$$\circ X^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(f_{ij} - n\hat{\theta}_{i.}\hat{\theta}_{.j})^2}{n\hat{\theta}_{i.}\hat{\theta}_{.j}} \xrightarrow{D} \chi_{(a-1)(b-1)}^2.$$

$$* df = k - 1 - \dim \Omega = ab - 1 - [(a-1) + (b-1)] = (a-1)(b-1).$$

**Example 10.2.** We have a table:

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$
$X = N$	17	11	11	14
$X = C$	17	9	8	7
$X = S$	12	13	19	28

Under the null hypothesis of independence, the MLE's are given by

$$\hat{\theta}_{.1} = \frac{46}{166}, \hat{\theta}_{.2} = \frac{33}{166}, \dots$$

Then the estimated expected counts  $n\hat{\theta}_{i.}\hat{\theta}_{.j}$  are given by the following table.

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$
$X = N$	14.6867	10.5361	12.1325	15.6446
$X = C$	11.3614	8.1506	9.3855	12.1024
$X = S$	19.9518	14.3133	16.4819	21.2530

Thus,  $T = X^2 = \frac{(17-14.6867)^2}{14.6867} + \dots + \frac{(28-21.2530)^2}{21.2530} = 11.7223$  and  $df = (a-1)(b-1) = 6$ . Thus  $p$ -value is 0.0685. Therefore, at 5% significance level, we do not have enough evidence to conclude that  $X$  and  $Y$  are dependent.

The code is:

```
1 chisq.test(rbind(c(17,11,11,14), c(17,9,8,7), c(12,13,19,28)))
```

### 10.2.2 $\chi^2$ test of homogeneity ( $X$ is deterministic)

- **Homogeneity** means the distributions of  $Y$  calculated for different category of  $X$  are all homogeneous, i.e., fixing the total number of each category of  $X$  in advance.  $X$  is not random anymore.
- Notation:
  - Let  $n_i$  be the marginal total of  $X = i$  category. We have  $\sum_i n_i = n$ . Marginal totals of all categories of  $X$  are fixed beforehand.
  - Instead of joint probabilities, we have bunch of conditional probabilities. Let  $\theta_{j|X=i} = P(Y = j|X = i)$ .
- We want to test  $H_0 : \theta_{j|X=1} = \theta_{j|X=2} = \cdots = \theta_{j|X=a} = \theta_j$ .
- Test statistic and corresponding distribution:
  - MLE of  $\theta_j$  will be  $\hat{\theta}_j = \sum_{i=1}^a \frac{f_{ij}}{n}$ .
  - $X^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(f_{ij} - n_i \hat{\theta}_j)^2}{n_i \hat{\theta}_j} \xrightarrow{D} \chi_{(a-1)(b-1)}^2$ .

**Example 10.3.** Of 279 participants in the study, 140 received a placebo and 139 received vitamin C. We have a table:

	No cold	Cold
Placebo	31	109
Vitamin C	17	122

Under the null hypothesis of independence, the MLE's are given by

$$\hat{\theta}_1 = \frac{48}{279} = 0.1720, \hat{\theta}_2 = \frac{231}{279} = 0.8280.$$

Then the estimated expected counts  $n_i \hat{\theta}_j$  are given by the following table.

	No cold	Cold
Placebo	24.08	115.92
Vitamin C	23.908	115.092

Thus,  $T = X^2 = \frac{(31-24.08)^2}{24.08} + \frac{(109-115.92)^2}{115.92} + \frac{(17-23.908)^2}{23.908} + \frac{(122-115.092)^2}{115.092} = 4.8124$  and  $df = (a-1)(b-1) = 1$ . Thus  $p$ -value is 0.0283. Therefore, at 5% significance level, we will reject the null hypothesis, i.e., there is relationship between taking vitamin C and the incidence of the common cold.

**Remark.** For  $\chi^2$  test of independence and homogeneity, we have an easy calculation of expected counts:

$$E_{ij} = \frac{i\text{th row total} \cdot j\text{th column total}}{\text{Grand total}}.$$

## 11 Correlation Coefficient and Least Square Regression

### 11.1 Relation among quantitative variables

- Suppose we have quantitative variables  $X$  and  $Y$ . Let  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$  be two corresponding data vectors. A visual display of these two vectors can be done by drawing a **scatter plot** that suggests the direction and magnitude of **correlation** between  $X$  and  $Y$ .
- **Pearson correlation coefficient** ( $r$ ) measures the linear relationship between two variables, where  $r \in [-1, 1]$ . If  $r = -1$ , it is perfect negative correlation. If  $r = 1$ , it is perfect positive correlation. If  $r = 0$ , it is zero correlation.
  - **Geometric definition** of  $r$  :  $r = \cos \theta$ , where  $\theta$  is the angle between  $n$  dimensional vector  $X$  and  $Y$ . Note:  $X$  and  $Y$  has to be centered.

### 11.2 Least square regression

- Let  $\hat{y} = b_1 + b_2x$  is the equation of the hypothetical line that we thought is going through the points, then  $(y_i - b_1 - b_2x_i)$  is the deviation of  $y_i$  from the line.
- Least square regression is finding the line that minimizes sum of the squared deviations:

$$\sum_{i=1}^n (y_i - b_1 - b_2x_i)^2.$$

- For  $b_1$ : Let

$$\frac{\partial}{\partial b_1} \sum (y_i - b_1 - b_2x_i)^2 = -2 \sum (y_i - b_1 - b_2x_i) = 0.$$

Then we have

$$\sum y_i - nb_1 - b_2 \sum x_i = 0 \Rightarrow b_1 = \bar{y} - b_2\bar{x}.$$

- For  $b_2$ : Let

$$\frac{\partial}{\partial b_2} \sum (y_i - b_1 - b_2x_i)^2 = 0.$$

Then we have

$$\begin{aligned}
\sum (y_i - b_1 - b_2 x_i) x_i &= \sum (x_i y_i - b_1 x_i - b_2 x_i^2) \\
&= \sum x_i y_i - (\bar{y} - b_2 \bar{x}) n \bar{x} - b_2 \sum x_i^2 \\
&= \sum x_i y_i - n \bar{x} \bar{y} + b_2 n \bar{x}^2 - b_2 \sum x_i^2 \\
&= \sum (x_i - \bar{x})(y_i - \bar{y}) - b_2 \sum (x_i - \bar{x})^2 = 0,
\end{aligned}$$

i.e.,

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

### 11.3 Classical linear regression under Normal distribution

- Assumptions:
  - $(Y_i | X_i = x_i) \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$ .
  - $Y_i$ 's are independent.
- Likelihood function:

$$L(\beta_1, \beta_2, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 \right].$$

For any given  $\sigma^2$ ,  $L(\beta_1, \beta_2, \sigma^2)$  will be maximized when residual sum of squares are minimized. Therefore,

$$\begin{aligned}
\hat{\beta}_1 &= b_1 = \bar{y} - b_2 \bar{x}, \\
\hat{\beta}_2 &= b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

- Interpretation of regression parameters:  $\beta_1$  represents the expected value of  $Y$  when  $X = 0$  and  $\beta_2$  represents the change in expected value of  $Y$  for one unit increase in  $X$ .

### 11.3.1 Properties of estimators of regression parameters

**Property 11.1.** Suppose  $Y$  is r.v., and  $x$  is treated as fixed constant, then we have

$$\begin{aligned} B_1 &= \bar{Y} - B_2 \bar{x}, \\ B_2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\sum_{i=1}^n (x_i - \bar{x})\bar{Y}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

Thus,  $B_1, B_2$  is a linear combination of  $Y_i$ 's, and both follow Normal distribution.

**Property 11.2.**  $B_1$  and  $B_2$  are unbiased estimators of  $\beta_1$  and  $\beta_2$ .

*Proof.* We have

$$\mathbb{E}[B_2] = \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$$

where

$$\begin{aligned} \mathbb{E}[(Y_i - \bar{Y})] &= \beta_1 + \beta_2 x_i - \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \beta_1 + \beta_2 x_i - \frac{1}{n} \sum_{i=1}^n (\beta_1 + \beta_2 x_i) \\ &= \beta_2 (x_i - \bar{x}). \end{aligned}$$

Thus

$$\mathbb{E}[B_2] = \frac{\sum_{i=1}^n (x_i - \bar{x})\beta_2(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_2,$$

and

$$\mathbb{E}[B_1] = \mathbb{E}[\bar{Y} - B_2 \bar{x}] = \beta_1.$$

□

**Property 11.3.**  $\text{Var}[B_2] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ .

*Proof.* We have  $B_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$ , then

$$\begin{aligned} \text{Var}[B_2] &= \text{Var}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \frac{1}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \sum_{i=1}^n \text{Var}[(x_i - \bar{x})Y_i] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}[Y_i]}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

□

### 11.3.2 Confidence interval and $t$ -test for $\beta_2$

- The unbiased estimator of  $\sigma^2$  is

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2,$$

and thus

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi_{(n-2)}^2.$$

Then we have

$$\frac{B_2 - \beta_2}{\sqrt{\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \bigg/ \sqrt{\frac{S^2}{\sigma^2}} = \frac{B_2 - \beta_2}{\sqrt{S^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{(n-2)}.$$

Therefore, the  $\gamma$  level CI for  $\beta_2$  is

$$\left[ B_2 \pm t_{\frac{1+\gamma}{2}(n-2)} \sqrt{\frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$

- Now we can test  $H_0 : \beta_2 = 0$ , i.e., there is no relationship between  $X$  and  $Y$ .



### 11.3.3 Sum of squares decomposition and ANOVA test

- Sum of squares decomposition:

- **Total sum of square**  $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ .

- $TSS$  can be written as the sum of two terms:

- (1) **Regression sum of square**  $RSS = b_2^2 \sum_{i=1}^n (x_i - \bar{x})^2$ .

- (2) **Error/Residual sum of square**  $ESS = \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$ .

- Coefficient of determination and correlation coefficient:

- Coefficient of determination is defined as  $R^2 = \frac{RSS}{TSS}$ .  $R^2$  is the proportion of variation in  $Y$  that can be explained by the model. For simple linear regression,  $r^2 = R^2$ .

**Example 11.1.** If  $R^2 = 0.98$ , then 98% variation in  $Y$  can be explained by the model.

- ANOVA table: another way of testing  $H_0 : \beta_2 = 0$ .

Source	df	Sum of Square (SS)	Mean SS = SS/df
X	1	$b_2^2 \sum (x_i - \bar{x})^2$	$b_2^2 \sum (x_i - \bar{x})^2$
Error	$n - 2$	$\sum (y_i - b_1 - b_2 x_i)^2$	$s^2$
Total	$n - 1$	$\sum (y_i - \bar{y})^2$	-

Therefore,  $F = \frac{\text{Mean SS for } X}{\text{Mean SS for error}} \sim F_{(1, n-2)}$ . The code for  $p$ -value is:

```
1 1 - pf(F, df.1, df.2)
```

### 11.3.4 Prediction and residual check

- Prediction: For  $X = x$ , predicted  $Y$  will be  $\hat{y} = b_1 + b_2 x$ . If the value  $x$  is within the range of the observed values of  $X$ , this prediction is called **interpolation**. If the value  $x$  is OUTSIDE the range, this prediction is called **extrapolation**.

- Residuals: For all observed  $X$  values we calculate predicted value  $\hat{y}$ . Residual corresponding to  $i$ th observation is  $(y_i - \hat{y}_i)$ . A positive residual indicates an ***under-prediction*** and a negative residual indicates an ***over-prediction***.

- Standardized residual plots: check model assumptions.

(1) Plot of standardized residuals against observed  $X$  values: the points should be clustered around zero (checking zero mean of the residuals) and look random (checking if the residuals are independent of  $X$  or not).

(2) Normal probability plot of the standardized residuals: the points should lie on a 45-degree line (checking the normality assumption).

The code for a simple linear model:

```

1 x = c(...)
2 y = c(...)
3 m = lm(y~x)
4 summary(m)
5 anova(m)
6 qqnorm(m$residuals)

```

## 11.4 Quantitative $Y$ and categorical $X$

- ***Dummy variable*** is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

**Example 11.2.** Let  $X_m = 1$  if male and  $X_m = 0$  if female. Let  $X_f = 1$  if female and  $X_f = 0$  if male. We have:

$Y$	Sex( $X$ )	$X_m$	$X_f$
10	Male	1	0
12	Male	1	0
8	Female	0	1
9	Female	0	1
...	...	...	...

From the Normality assumption, we have  $Y|X \sim \mathcal{N}(\beta_1 X_m + \beta_2 X_f, \sigma^2)$ , and hence  $\mathbb{E}[Y|X] = \beta_1 X_m + \beta_2 X_f$ ,  $\mathbb{E}[Y|X = \text{Male}] = \beta_1$ ,  $\mathbb{E}[Y|X =$

Female] =  $\beta_2$ . So  $\beta_1$  is the population mean of  $Y$  for male and  $\beta_2$  is the population mean of  $Y$  for female.

A common hypothesis that is tested is  $H_0 : \beta_1 = \beta_2 \Rightarrow \beta_1 - \beta_2 = 0$ , i.e., there is no difference in mean of  $Y$  between male and female, or there is no relationship between  $X$  and  $Y$ .

Another method: Assume  $Y|X \sim \mathcal{N}(\beta_1 + \beta_2 X_f, \sigma^2)$ , then  $\mathbb{E}[Y|X = \text{Male}] = \beta_1$ ,  $\mathbb{E}[Y|X = \text{Female}] = \beta_1 + \beta_2$ .

Thus,  $\mathbb{E}[Y|X = \text{Female}] - \mathbb{E}[Y|X = \text{Male}] = \beta_2$ , and  $\beta_2$  gives us the difference mean of the two groups. If  $\beta_2 = 0$ , there is no difference between the groups, i.e.,  $X$  and  $Y$  are not related.

- If  $X$  has more than two categories, number of dummy variables needed and number of corresponding  $\beta$ 's will increase. We will compare the means by comparing two of them at a time, which is called ***multiple comparisons***.