STATISTICAL INFERENCE

POPULATION AND SAMPLE

\[
\begin{align*}
\text{Population} & = \text{all elements of interest} \\
& \text{Characterized by a distribution } F \\
& \text{with some parameter } \theta \\
\downarrow \\
\text{Sample} & = \text{the data } X_1, \ldots, X_n, \\
& \text{selected subset of the population}
\end{align*}
\]

\[ n = \text{sample size} \]

Examples of \( F \): Bernoulli(\( p \)), Normal(\( \mu, \sigma \)), Gamma(\( n, \lambda \)), Poisson(\( \lambda \)), etc.
Statistical Inference

Statistical Inference = inference about the population based on a sample

- Parameter estimation
- Confidence intervals
- Hypothesis testing
- Model fitting
Parameter Estimation

**Statistic** = any function of data \( W(X_1, ..., X_n) \)

**Estimator of** \( \theta \) = any statistic used to estimate parameter \( \theta \)

Estimator \( \hat{\theta} \) is **unbiased** if \( \mathbb{E}(\hat{\theta}) = \theta \)

**Standard error** of an estimator is its standard deviation \( \text{Std}(\hat{\theta}) \)

It is estimated by \( \text{Std}(\hat{\theta}) \). It shows the accuracy, reliability of estimator \( \hat{\theta} \).
Estimation of a mean

Sample \((X_1, \ldots, X_n)\) is collected from a population with \(E(X) = \mu\) and \(\text{Var}(X) = \sigma^2\).

Estimate the population mean \(\theta = \mu = E(X_i)\) by a sample mean

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

Properties:

\[
E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \theta = \theta
\]

\[
\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\]

So, \(\bar{X}\) is unbiased, and its standard error is

\[
SE(\bar{X}) = \text{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}
\]
Estimation of a variance

Estimate the population variance

\[ \theta = \sigma^2 = \text{Var}(X_i) \]

by a sample variance

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

It is also unbiased: \( E(S^2) = \sigma^2 \).

Then, the standard error of \( \bar{X} \) is estimated by

\[ \text{Std}(\bar{X}) = \frac{S}{\sqrt{n}} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n(n-1)}} \]
Estimation of a proportion

Sample \((X_1, \ldots, X_n)\) is collected from Bernoulli population with parameter \(p\).

Estimate the population proportion \(p = E(X_i)\) by a sample proportion

\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{\text{number of } X_i = 1}{n}
\]

Special case of a sample mean \(\bar{X}\)

\[
E(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}
\]

So, \(\hat{p}\) is unbiased;

its standard error is \(SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}\)

typically estimated by \(\hat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\)
General Methods of Estimation

1. Method of Moments

\[ \mu_k = \text{E} X^k \]
\[ M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \]

To estimate \( d \) parameters, solve the system of \( d \) equations

\[
\begin{cases}
M_1 = \mu_1 \\
... \\
M_d = \mu_d
\end{cases}
\]

\( M_1, \ldots, M_d \) are known from the sample; \( \mu_1, \ldots, \mu_d \) are functions of unknown parameters
Example: $X_1, \ldots, X_n$ are Exponential($\lambda$)

Estimate $\lambda$.

The number of parameters is $d = 1$. So, we need 1 equation.

$M_1 = \bar{X}; \; \mu_1 = \frac{1}{\lambda}$

Solve

$M_1 = \mu_1 \Rightarrow \bar{X} = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{\text{mom}} = \frac{1}{\bar{X}}$
2. Method of Maximum Likelihood

Maximize the probability (pmf, pdf) of seeing the really observed data

**Implementation**

Observe $X_1, \ldots, X_n$ from pdf or pmf $f(x \mid \theta)$. 

Maximize

$$f(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

in $\theta$.

Simplification: maximize

$$\ln f(X_1, \ldots, X_n \mid \theta) = \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

Typically, compute

$$\frac{\partial}{\partial \theta} \ln f(X_1, \ldots, X_n \mid \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i \mid \theta)$$

equate to 0 and solve in $\theta$. 

Method of maximum likelihood

Example: $X_1, \ldots, X_n$ are Exponential($\lambda$)

$$f(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$

$$\ln f(X_1, \ldots, X_n \mid \theta) = \sum_{i=1}^{n} \ln \left( \lambda e^{-\lambda X_i} \right)$$

$$= n \ln \lambda - \lambda \sum_{i=1}^{n} X_i$$

$$\frac{\partial}{\partial \lambda} \ln f(X_1, \ldots, X_n \mid \theta) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i =: 0$$

Solve for $\lambda$,

$$\hat{\lambda}_{mle} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}$$
Maximum likelihood

This area

\[ P \{ x - h \leq x \leq x + h \} \approx (2h) f(x) \]

Probability of observing “almost” \( X = x \)
Confidence Intervals

A **100 \((1-\alpha)\) % confidence interval** is an interval that contains parameter \(\theta\) with probability \(\gamma\).

That is,

\[
P\{a \leq \theta \leq b\} = 1 - \alpha
\]

where

\[
a = a(X_1, \ldots, X_n) \quad \text{and} \quad b = b(X_1, \ldots, X_n)
\]

are statistics. So, \(a\) and \(b\) are random, \(\theta\) is not.
Confidence intervals for the same parameter $\theta$ obtained from different samples of data.

Confidence intervals and coverage of parameter $\theta$. 
Example: $X_1, \ldots, X_n$ from $\text{Normal}(\mu, \sigma)$ with unknown $\mu$, known $\sigma$

1. Estimate $\theta = \mu$ by its estimator $\bar{X} = \frac{1}{n} \sum X_i$.

2. Find its distribution: $\text{Normal}$ with

$$\mathbb{E}(\bar{X}) = \mu$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{1}^{n} \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Therefore,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is Normal}(0,1)$$

3. Find critical values $\pm z_{\alpha/2}$ such that

$$P\left\{-z_{\alpha/2} < Z < z_{\alpha/2}\right\}$$

for $Z \sim \text{Normal}(0,1)$. 

14
4. Then we have

\[ P \left\{ -\frac{z_{\alpha/2}}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{z_{\alpha/2}}{\sigma/\sqrt{n}} \right\} = 1 - \alpha \]

Solve for \( \mu \):

\[ P \left\{ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right\} = 1 - \alpha \]

5. Hence,

\[ \bar{X} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \left[ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right] \]

is a \((1 - \alpha)100\%\) confidence interval for \( \mu \).

\( \bar{X} \) is approximately Normal for large \( n \) and any distribution of \( X_1, \ldots, X_n \).
**Confidence intervals**

**When \( \sigma \) is unknown**

Data \( X_1, ..., X_n \) from Normal(\( \mu, \sigma \)) with unknown \( \mu \), **unknown** \( \sigma \)

1. Estimate \( \sigma \) by 
\[
S = \sqrt{\frac{1}{n - 1} \sum_{1}^{n} (X_i - \bar{X})^2}
\]

2. Use \( t \)-distribution with \((n - 1)\) degrees of freedom instead of Normal.

For large \( n \), use Normal approximation.

**Result:**
\[
\bar{X} \pm \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}
\]
Hypothesis $H_0$ and alternative $H_A = \text{mutually exclusive statements about the unknown parameter } \theta$.

Collect data
$$\downarrow$$
Conduct a test
$$\downarrow$$
State if there is sufficient evidence to reject $H_0$ in favour of $H_A$.

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>Reject $H_0$</th>
<th>Accept $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>Type I error</td>
<td>correct</td>
</tr>
<tr>
<td>$H_0$ is false</td>
<td>correct</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

Control the **significance level**

$$\alpha = P \{ \text{Type I error} \}$$
Hypotheses testing

Data: $X_1, \ldots, X_n$ from $\text{Normal}(\mu, \sigma)$ with unknown $\mu$, known $\sigma$

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.

1. Find $\pm z_{\alpha/2}$. Acceptance region: $[-z_{\alpha/2}, z_{\alpha/2}]$.

2. Compute the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$ 

3. If $Z$ belongs to the acceptance region, do not reject $H_0$. Otherwise, reject $H_0$.

If $H_0$ is true, $Z$ has $\text{Normal}(0,1)$ distribution, and

$$P \{ \text{Type I error } \} = P \{|Z| > z_{\alpha/2}\} = \alpha$$
Hypotheses testing

One-sided, right-tail tests

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$.

1. Find $z_\alpha$. The acceptance region is $(-\infty, z_\alpha]$.

2. Compute the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$ 

3. If $Z$ belongs to the acceptance region, do not reject $H_0$.
   Otherwise, reject $H_0$. 

19
One-sided, left-tail tests

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu < \mu_0$.

1. Find $z_\alpha$. The acceptance region is $[-z_\alpha, +\infty)$.

2. Compute the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}.$$ 

3. If $Z$ belongs to the acceptance region, do not reject $H_0$.
Otherwise, reject $H_0$. 

20
Hypotheses testing

Case of unknown variance

1. Estimate $\sigma$ by $S = \sqrt{\frac{1}{n-1} \sum_{1}^{n} (X_i - \bar{X})^2}$

2. Use $t$-distribution with $(n - 1)$ degrees of freedom.

For large $n$, use Normal approximation.
### Hypotheses testing, Z-tests

**Null hypothesis**

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>Parameter, estimator</th>
<th>If $H_0$ is true:</th>
<th>Test statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$, $\hat{\theta}$</td>
<td>$E(\hat{\theta})$, $\text{Var}(\hat{\theta})$</td>
<td>$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}}$</td>
<td></td>
</tr>
</tbody>
</table>

**One-sample Z-tests for means and proportions, based on a sample of size $n$**

| $\mu = \mu_0$ | $\mu$, $\bar{X}$ | $\mu_0$ | $\frac{\sigma^2}{n}$ | $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ |
| $p = p_0$ | $p$, $\hat{p}$ | $p_0$ | $\frac{p_0(1 - p_0)}{n}$ | $\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})}}$ |

**Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size $n$ and $m$**

| $\mu_X - \mu_Y = D$ | $\mu_X - \mu_Y$, $\bar{X} - \bar{Y}$ | $D$ | $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ | $\frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$ |
| $p_1 - p_2 = D$ | $p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$ | $D$ | $\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$ | $\frac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}}$ |
**Hypothesis testing, t-tests**

<table>
<thead>
<tr>
<th>Hypothesis $H_0$</th>
<th>Conditions</th>
<th>Test statistic $t$</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = \mu_0$</td>
<td>Sample size $n$; unknown $\sigma$</td>
<td>$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$</td>
<td>$n - 1$</td>
</tr>
<tr>
<td>$\mu_X - \mu_Y = D$</td>
<td>Sample sizes $n, m$; unknown but equal $\sigma_X = \sigma_Y$</td>
<td>$t = \frac{\bar{X} - \bar{Y} - D}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$</td>
<td>$n + m - 2$</td>
</tr>
<tr>
<td>$\mu_X - \mu_Y = D$</td>
<td>Sample sizes $n, m$; unknown, unequal $\sigma_X \neq \sigma_Y$</td>
<td>$t = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$</td>
<td>Special formula</td>
</tr>
</tbody>
</table>