Chapter 5 Statistical Inference

5.2 Probability Model

\( \text{\textcircled{96}} \): Life span of an iPhone \( \sim \) exponential(1)

95% interval \((0, c)\) where

\[
0.95 = \int_0^c e^{-x} \, dx = 1 - e^{-c} \Rightarrow c = -\log(0.05) = 2.9957
\]

A range, likely in most cases, for the life span of an iPhone.

An educated guess.

5.3 Statistical Model

\( \Theta \) (parameter of the model) \( \subseteq \Omega \) (parameter space)

Model \( \Theta_1 = M \Theta_2 \iff \Theta_1 = \Theta_2 \)

Assume \( X \) has density \( f_\Theta \), take a simple random sample \( X_1, X_2, \ldots, X_n \) (identically distributed and independent \( i.i.d. \)), their joint density is \( f_{\Theta_1}(X_1) f_{\Theta_2}(X_2) \cdots f_{\Theta_n}(X_n) \), which is a model for a sample.

\( \text{\textcircled{96}} \): Life span \( X_1 \sim X_5 : (5, 3.5, 3.3, 4.1, 2.8) \) exponential(1) \( \Theta ?? \)

\( X^* \sim X^* : (2, 2.5, 1.8, 1.1, 3.2) \) \( \Theta ?? \)

different values. exponential(1) is model assumption. Model checking needed.
5.4. Data Collection
Observational : passive., randomness not guaranteed
Population: all the objects.
Finite Population: finite # of objects
Quantitative variable: numerical values. \( X \) (age)
Categorical variable: gender
different analysis tools

Population distribution = Probability distribution (model)
\( F(x) = P(X \leq x) \).

Empirical distribution estimated from data
\[ \hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i). \]
\# of samples \( n \) \leq X

Sampling: 1 simple random sampling
draw objects without replacement
2 representative sample (IRS), tails of income distribution
online sample — selection bias.

Sample size requirement:
very small sample size \( \Rightarrow \) no reliable conclusion.
Survey sample: non-response rate.
Basic graphical inferences:

1. **Histogram** - numerical (quantitative) variable. A sample $X_1, \ldots, X_n$, divided into intervals $[h_1, h_2], [h_2, h_3], \ldots, [h_{m-1}, h_m]$. Equal sized.

2. **Box plot**: $Q_3 - Q_1 = \text{interquartile range.}$

3. **QQ plot** (normal) - numerical variable.

- $X(1) \leq X(2) \leq \cdots \leq X(n)$
- vs.
- $Y(1) \leq Y(2) \leq \cdots \leq Y(n)$

$\widetilde{Y}(p) \leq \widetilde{Y}(p_2) \leq \cdots \leq \widetilde{Y}(p_n)$

- $p_i$ plotting positions, (probabilities)

- e.g. $p_i = \frac{i - 0.5}{n}$
5.4.20. A population \( \Phi \) contains 2 subpopulations \( \Phi_1 \) and \( \Phi_2 \). Size of \( \Phi_1 \) is \( p \) (0 ≤ \( p \) ≤ 1). Measure a variable \( X \) on a selected object \( \xi \) from \( \Phi_1 \).

1. \( f_1(x) = p f_1(x) + (1-p) f_2(x) \).
2. \( M = p M_1 + (1-p) M_2 = \int_{-\infty}^{\infty} x (pf_1(x) + (1-p)f_2(x)) \, dx \).
3. \( \sigma^2 = \left( \int_{-\infty}^{\infty} x^2 [pf_1(x) + (1-p)f_2(x)] \, dx \right) - M^2 \)
   \[\sigma^2 = p(\sigma_2^2 + M_2^2) + (1-p)(\sigma_1^2 + M_1^2) - (pM_1 + (1-p)M_2)^2 \]
   \[\sigma^2 = p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(M_1 - M_2)^2 \].
4. Take i.i.d. sample of \( N_1 = pn \) and \( N_2 = (1-p)n \)
   from \( \Phi_1 \) and \( \Phi_2 \) respectively.

Proportional stratified sampling,
\[ \bar{X}_1 = \hat{M}_1, \quad \bar{X}_2 = \hat{M}_2, \quad p \bar{X}_1 + (1-p) \bar{X}_2 = \hat{M}. \]
Descriptive Statistics:
a sample $x_1, x_2, \ldots, x_n \sim \text{i.i.d. } f(x)$
sample mean $\overline{X} = (x_1 + x_2 + \ldots + x_n) / n$
sample variance $s^2 = \frac{1}{n-1} \sum (x_i - \overline{X})^2$
sample standard deviation $\sqrt{s^2} = sd.$
pth quantile ($0 \leq p \leq 1$) $t_p$ defined as $F(t_p) = p.$
sample quantiles:
ordered $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
we have $\frac{i-1}{n} < p \leq \frac{i}{n}$.

$p \hat{=} t_p = x_{(i)} + n(\overline{x} - x_{(i-1)}) \times (p - \frac{i-1}{n})$
linear interpolation b/w $x_{(i)}$ and $x_{(i-1)}$.

$p = 0.25, 0.5, 0.75, 0.95$
Q1, Q2, Q3 quartiles.

$p = 0.5$ for median (a robust center of dist'n).

Statistical Inference:
1. estimate a parameter - point estimation
2. confidence interval/region for $\theta$
3. hypothesis test (parametric $\theta$ value)
4. bayes analysis.
5.5.2 Waiting time (minutes)

0, 1min 2m 3m 4m 5m 10m 15m

1 2 3 4 5 8

5.5.2 Waiting time (minute)

15, 10, 2, 3, 1, 0, 4, 5, 5, 3, 3, 4, 2, 1, 4, 5

0m 1m 2m 3m 4m 5m 10m 15m

1 \underline{2} \underline{3} 3 3 3 1 1. total = 16

(b). density for proportion

\[ \begin{array}{c}
\frac{3}{16} \\
\frac{2}{16} \\
\frac{1}{16} \\
\frac{1}{16}
\end{array} \]

(a). Empirical distribution

\[ \begin{array}{c}
\frac{3}{16} \\
\frac{3}{16} \\
\frac{1}{16}
\end{array} \]

(b) \( \bar{X} = \frac{\text{sum}}{16} = 4.188 \) \( \bar{S}^2 = 13.63 \)

(c) Q1 = 2, Q2 = 3.5, Q3 = 5. IQR = 3