Statistical Methods for Data Science
Lessons 26 - Confidence intervals: mean, proportion, linear regression

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it
From point estimate to interval estimate

**Estimator and point estimate**

A *statistics* is a function of \( h(X_1, \ldots, X_n) \) of r.v.'s.

An *estimator* of a parameter \( \theta \) is a statistics \( T_n = h(X_1, \ldots, X_n) \) intended to provide information about \( \theta \).

A *point estimate* \( t \) of \( \theta \) is \( t = h(x_1, \ldots, x_n) \) over realizations of \( X_1, \ldots, X_n \).

- Sometimes, a *range* of plausible values for an unknown parameter is preferred
- Idea: *confidence interval* is an interval for which we can be confident the unknown parameter is in with a specified probability (*confidence level*)
Example

• From the Chebyshev’s inequality:

\[ P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \]

For \( Y = \bar{X}_n \), \( k = 2 \) and \( \sigma = 100 \text{ Km/s} \):

\[ P(|\bar{X}_n - \mu| < 200) \geq 0.75 \]

▶ i.e., \( \bar{X}_n \in (\mu - 200, \mu + 200) \) with probability \( \geq 75\% \)  

▶ or, \( \mu \in (\bar{X}_n - 200, \bar{X}_n + 200) \) with probability \( \geq 75\% \)

• \( (\bar{X}_n - 200, \bar{X}_n + 200) \) is an interval estimator of the unknown \( \mu \)
  ▶ the interval contains \( \mu \) with probability \( \geq 75\% \)

• Let \( t = 299852.4 \) be the point estimate (realization of \( T = \bar{X}_n \))
• \( \mu \in (t - 200, t + 200) = (299652.4, 300052.4) \) is correct with confidence \( \geq 75\% \)
The better the estimator the better (smaller) the intervals

- Assume $X_i \sim N(\mu, \sigma^2)$. Hence, $\bar{X}_n \sim N(\mu, \sigma^2/n)$ and:

$$Z_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1)$$

- $P(|Z_n| \leq 1.15) = P(-1.15 \leq Z_n \leq 1.15) = \Phi(1.15) - \Phi(-1.15) = 0.75$
  
  - $-1.15 = q_{0.125}$ and $1.15 = q_{0.875}$ are called the critical values for achieving 75% probability

- Going back to $\bar{X}_n$:

$$P(-1.15 \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq 1.15) = P(\bar{X}_n - 1.15 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.15 \frac{\sigma}{\sqrt{n}}) = 0.75$$

- $\mu \in (t - 1.15 \frac{200}{\sqrt{100}}, t + 1.15 \frac{200}{\sqrt{100}}) = (t - 23, t + 23)$ is correct with confidence $= 75\%$
Confidence intervals

Suppose a dataset $x_1, \ldots, x_n$ is given, modeled as realization of random variables $X_1, \ldots, X_n$. Let $\theta$ be the parameter of interest, and $\gamma$ a number between 0 and 1. If there exist sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that

$$P(L_n < \theta < U_n) = \gamma$$

for every value of $\theta$, then $(l_n, u_n)$, where $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$, is called a $100\gamma\%$ confidence interval for $\theta$. The number $\gamma$ is called the confidence level.

- Sometimes, only have $P(L_n < \theta < U_n) \geq \gamma$ [conservative $100\gamma\%$ confidence interval]
  - E.g., the interval found using Chebyshev’s inequality
- There is no way of knowing if $l_n < \theta < u_n$ (interval is correct or not)
- We only know that we have probability $\gamma$ of covering $\theta$
- Notation: $\gamma = 1 - \alpha$ where $\alpha$ is called the significance level
  - $100\gamma = 95\%$ confidence level, i.e. probability that interval includes the parameter
  - $\alpha = 0.05$ significance level, i.e. probability that interval does not include the parameter

Seeing theory simulation
Confidence interval for the mean

• Let $X_1, \ldots, X_n$ be a random sample and $\mu = E[X_i]$ to be estimated

• Problem: confidence intervals for $\mu$?
  ▶ Normal data
    □ with known variance
    □ with unknown variance
  ▶ General data (with unknown variance)
    □ large sample, i.e., large $n$
    □ bootstrap (next lesson)
Critical values

The (right) critical value $z_p$ of $Z \sim N(0, 1)$ is the number with right tail probability $p$:

$$P(Z \geq z_p) = p$$

- Alternatively, $p = 1 - \Phi(z_p) = 1 - P(Z \leq z_p)$.
  - This is why Table B.1 of the textbook is given for $1 - \Phi()$

- Alternatively, $\Phi(z_p) = 1 - p$, i.e., $z_p$ is the $(1 - p)$th quantile

- Since $P(Z \geq z_p) = P(Z \leq -z_p) = p$, we have:
  $$P(Z \geq -z_p) = 1 - P(Z \leq -z_p) = 1 - p$$

  and then:
  $$z_{1-p} = -z_p$$

  - E.g., $z_{0.975} = -z_{0.025} = -1.96$ and $z_{0.025} = -z_{0.975} = 1.96$
CI for the mean: normal data with known variance

- Dataset \( x_1, \ldots, x_n \) realization of random sample \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \)
- Estimator \( \bar{X}_n \sim N(\mu, \sigma^2/n) \) and the scaled mean:

\[
Z = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1)
\]  

(1)

- Confidence interval for \( Z \):

\[
P(c_l \leq Z \leq c_u) = \gamma \quad \text{or} \quad P(Z \leq c_l) + P(Z \geq c_u) = \alpha = 1 - \gamma
\]

- Symmetric split:

\[
P(Z \leq c_l) = P(Z \geq c_u) = \alpha/2
\]

Hence \( c_u = -c_l = z_{\alpha/2} \), and by (1):

\[
P(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha = \gamma
\]

\((\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})\) is a 100\(\gamma\)% or 100(1 - \(\alpha\))% confidence interval for \( \mu \)
One-sided confidence intervals

• One-sided confidence intervals (*greater-than*):

\[ P(L_n < \theta) = \gamma \]

Then \((l_n, \infty)\) is a \(100\gamma\%\) or \(100(1 - \alpha)\%\) one-sided confidence interval

• \(l_n\) is called the lower confidence bound

• Normal data with known variance:

\[ P(\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu) = 1 - \alpha = \gamma \]

\((\bar{x}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)\) is a \(100\gamma\%\) or \(100(1 - \alpha)\%\) one-sided confidence interval for \(\mu\)

See R script
CI for the mean: normal data with unknown variance

- Use the unbiased estimator of $\sigma^2$ and its estimate

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

- and then $S^2_n / n$ is an unbiased estimator of $Var(\bar{X}_n) = \sigma^2 / n$

- The following transformation is called the studentized mean: $T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim t(n-1)$

- Student t-distribution $X \sim t(m)$:
  - for $m = 1$, it is the Cauchy distribution
  - $E[X] = 0$ for $m \geq 2$, and $Var(X) = m/(m-2)$ for $m \geq 3$
  - For $m \to \infty$, $X \to N(0,1)$

---

**Definition.** A continuous random variable has a $t$-distribution with parameter $m$, where $m \geq 1$ is an integer, if its probability density is given by

$$f(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}$$

for $-\infty < x < \infty$,

where $k_m = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}}$. This distribution is denoted by $t(m)$ and is referred to as the $t$-distribution with $m$ degrees of freedom.

---

Some history on its discovery

See R script
CI for the mean: normal data with unknown variance

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

### Critical value

The (right) **critical value** $t_{m,p}$ of $T \sim t(m)$ is the number with right tail probability $p$:

$$P(T \geq t_{m,p}) = p$$

- Same properties of $z_p$
- From the studentized mean:

$$T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim t(n-1)$$

to confidence interval:

$$P(\bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}) = 1 - \alpha = \gamma$$

$$(\bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}})$$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for $\mu$

See R script
CI for the mean: general data with unknown variance

- Dataset \( x_1, \ldots, x_n \) realization of random sample \( X_1, \ldots, X_n \sim F \)
- A variant of CLT states that for \( n \to \infty \)

\[
T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \to N(0, 1)
\]

- For large \( n \), we make the approximation: [how large should \( n \) be?]

\[
T = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \approx N(0, 1)
\]

and then

\[
P(\bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}}) \approx 1 - \alpha = \gamma
\]

\((\bar{x}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}})\) is a 100\(\gamma\)% or 100(1 - \(\alpha\))% confidence interval for \(\mu\)

See R script
General form of Wald confidence intervals

\[ \theta \in \hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}) \quad \text{or} \quad \theta \in \hat{\theta} \pm t_{\alpha/2} se(\hat{\theta}) \]

- They originate from the **Wald test statistics**:

  \[ T = \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{se(\hat{\theta})} \]

- Importance of standard error \( se(\hat{\theta}) \) of estimators!
- Limitation: asymptotic, symmetric intervals
CI for proportions (e.g., classifier accuracy)

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim Ber(p)$
  - $x_i = \mathbb{1}_{y^+_\theta(w_i) = c_i}$ is 1 for correct classification, 0 for incorrect classification [over a test set]
  - $p$ is the (unknown) misclassification error of classifier $y^+_\theta()$

- $B = \sum_{i=1}^n X_i \sim Bin(n, p)$ and $b = \sum_{i=1}^n x_i$ (number of observed successes)
  - For small $n$, we can build exact bounds $(p_L, p_U)$ such that: [Exact interval]

\[
\begin{align*}
  l_B &= \min_{\theta} \left\{ \sum_{x=B}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} \geq \frac{\alpha}{2} \right\} \\
  u_B &= \max_{\theta} \left\{ \sum_{x=0}^{B} \binom{n}{x} \theta^x (1-\theta)^{n-x} \geq \frac{\alpha}{2} \right\}
\end{align*}
\]

- $l_B$ is the smallest $\theta$ for which $P(X \geq b) \geq \frac{\alpha}{2}$ for $X \sim Bin(n, \theta)$
- $u_B$ is the greatest $\theta$ for which $P(X \leq b) \geq \frac{\alpha}{2}$ for $X \sim Bin(n, \theta)$

\[
P(l_B \leq p \leq u_B) = 1 - \alpha
\]

and then $(l_b, u_b)$ is a $100\gamma\%$ or $100(1 - \alpha)\%$ confidence interval for $p$

- Closed form given using quantiles of the F-distribution
Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R

Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).
Cl for proportions (e.g., classifier accuracy)

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim \text{Ber}(p)$
  - $x_i = \mathbb{1}_{y_{\theta}^+(w_i) = c_i}$ is 1 for correct classification, 0 for incorrect classification
  - $p$ is the (unknown) misclassification error of classifier $y_{\theta}^+$

- $B = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$
  - For large $n$, $\text{Bin}(n, p) \approx N(np, np(1-p))$ for $0 \ll p \ll 1$ [De Moivre–Laplace]
    - $\text{se}(B) = \sqrt{n\bar{X}_n(1-\bar{X}_n)}$ because $B \sim \text{Bin}(n, p)$
    - We make the approximation: $T = (B - np)/\text{se}(B) \approx N(0,1)$ and then
      $$P(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq p \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}) \approx 1 - \alpha = \gamma$$
      $$\left(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}\right)$$ is a 100$\gamma$% or 100$(1 - \alpha)$% confidence interval for $p$
    - This is a Wald confidence interval!

- Drawbacks: symmetric, large sample, skewness, etc.
  - [Wilson score interval]

See R script
Determining the sample size

- For a fixed $\alpha$, the narrower the CI the better (smaller variability)
- Sometimes, we start with an accuracy requirement (maximal width $w$ of the interval):
  - find a $100(1 - \alpha)\%$ CI $(l_n, u_n)$ such that $u_n - l_n \leq w$

- How to set $n$ to satisfy the $w$ bound?
- Case: normal data with known variance $\sigma^2$
  - CI is $(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$
  - Bound on the CI is:
    \[ 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq w \]
    leading to:
    \[ n \geq \left(2z_{\alpha/2} \frac{\sigma}{w}\right)^2 \]

- Case $\sigma^2$ unknown: use estimate $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
Confidence intervals for simple linear regression coefficients

- Simple linear regression: \( Y_i = \alpha + \beta x_i + U_i \) with \( U_i \sim \mathcal{N}(0, \sigma^2) \) and \( i = 1, \ldots, n \)

- We have \( \hat{\beta} \sim \mathcal{N}(\beta, \text{Var}(\hat{\beta})) \) where \( \text{Var}(\hat{\beta}) = \sigma^2 / SXX \) is unknown [see Lesson 20]

- The Wald statistics is \( t(n - 2) \)-distributed: [proof omitted]

\[
\frac{\hat{\beta} - \beta}{\sqrt{\text{Var}(\hat{\beta})}} \sim t(n - 2)
\]

- For \( \gamma = 0.95 \):

\[
P\left(-t_{n-2,0.025} \leq \frac{\hat{\beta} - \beta}{\sqrt{\text{Var}(\hat{\beta})}} \leq t_{n-2,0.025}\right) = 0.95
\]

and then a 95% confidence interval is: \( \hat{\beta} \pm t_{n-2,0.025} \text{se}(\hat{\beta}) \) where \( \text{se}(\hat{\beta}) = \hat{\sigma} / \sqrt{SXX} \)

- Similarly, we get for \( \alpha \), \( \hat{\alpha} \pm t_{n-2,0.025} \text{se}(\hat{\alpha}) \)

See R script
Confidence intervals of fitted values

- Simple linear regression: $Y_i = \alpha + \beta x_i + U_i$ with $U_i \sim \mathcal{N}(0, \sigma^2)$ and $i = 1, \ldots, n$
- For the fitted values $\hat{y} = \hat{\alpha} + \hat{\beta} x_0$ at $x_0$, a 95% confidence interval is:

$$\hat{y} \pm t_{n-2,0.025} \text{se}(\hat{y})$$

where $\text{se}(\hat{y}) = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{S_{XX}}\right)}$ [see Lesson 21]
- This interval concerns the expectation of fitted values at $x_0$.
  - E.g., the mean of predicted values at $x_0$ is in $[\hat{y} + t_{n-2,0.025} \text{se}(\hat{y}), \hat{y} - t_{n-2,0.025} \text{se}(\hat{y})]$

See R script
Prediction intervals of fitted values

- Simple linear regression: $Y_i = \alpha + \beta x_i + U_i$ with $U_i \sim \mathcal{N}(0, \sigma^2)$ and $i = 1, \ldots, n$
- For a given single prediction, we must also account for the error term $U$ in:

$$\hat{V} = \hat{\alpha} + \hat{\beta}x_0 + U$$

- Assuming $U \sim \mathcal{N}(0, \sigma^2)$, we have $\text{Var}(\hat{V}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)$
- A 95% confidence interval is:

$$\hat{y} \pm t_{n-2,0.025} \text{se}(\hat{V})$$

where $\text{se}(\hat{V}) = \hat{\sigma} \sqrt{\left(1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX}\right)}$

- A predicted value at $x_0$ is in $[\hat{y} - t_{n-2,0.025} \text{se}(\hat{V})$ and $\hat{y} + t_{n-2,0.025} \text{se}(\hat{V})]$

See R script
• On confidence intervals and statistical tests (with R code)

Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)
Nonparametric Statistical Methods.
3rd edition, John Wiley & Sons, Inc.