Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lessons 26 - Confidence intervals: mean, proportion, linear regression

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
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 θ is a statistics $T_n = h(X_1, \ldots, X_n)$ intended to provide information about θ.
A *point estimate* t of θ is $t = h(x_1, \ldots, x_n)$ over realizations of $X_1, \ldots, X_n$.

- Sometimes, a *range* of plausible values $l < \theta < u$ is useful, as it provides uncertainty information
- Idea: *confidence interval* is an interval for which we can be confident the unknown parameter θ is in with a specified probability (called *confidence level*)

Example

Table 17.1. Michelson data on the speed of light.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

- From the Chebyshev's inequality:

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

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

$$P(|\bar{X}_n - \mu| < 200) \geq 1 - \frac{1}{2^2} = 0.75$$

  - i.e., $\bar{X}_n \in (\mu - 200, \mu + 200)$ with probability $\geq 75\%$ *[random variable in a fixed interval]*
  - or, $\mu \in (\bar{X}_n - 200, \bar{X}_n + 200)$ with probability $\geq 75\%$ *[fixed value in a random interval]*
- $(\bar{X}_n - 200, \bar{X}_n + 200)$ is an interval estimator of the unknown μ
  - the interval contains μ with probability $\geq 75\%$
- Let $\bar{x}_n = 299\,852.4$ be the point estimate (realization of $\bar{X}_n$)
- $\mu \in (\bar{x}_n - 200, \bar{x}_n + 200) = (299\,652.4, 300\,052.4)$ is correct *with confidence* $\geq 75\%$

The smaller the interval, the better the estimator

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

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

- $P(-1.15 \leq Z_n \leq 1.15) = \Phi(1.15) - \Phi(-1.15) = 0.75$
  - $-1.15 = q_{0.125}$ and $1.15 = q_{0.875}$ are called *the critical values* for achieving 75% probability
- Going back to $\bar{X}_n$:

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

- $\mu \in (\bar{x}_n - 1.15\frac{100}{\sqrt{100}}, \bar{x}_n + 1.15\frac{100}{\sqrt{100}}) = (\bar{x}_n - 11.5, \bar{x}_n + 11.5) = (299\,840.9, 299\,863.9)$ is correct *with confidence* $= 75\%$

Confidence intervals

CONFIDENCE INTERVALS. Suppose a dataset $x_1, \ldots, x_n$ is given, modeled as realization of random variables $X_1, \ldots, X_n$. Let θ be the parameter of interest, and γ 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

$$\mathrm{P}(L_n < \theta < U_n) = \gamma$$

for every value of θ, 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* θ. The number γ 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 γ of covering θ
- Notation: $\gamma = 1 - \alpha$ where α 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 intervals for the mean: summary

- $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n \sim F$ with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
- Problem: what is a confidence interval for μ?
  - Normal data $F = \mathcal{N}(\mu, \sigma^2)$
    - with known variance: $Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$
    - with unknown variance: $T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}}$
  - General data (with unknown variance)
    - large sample, i.e., large n: $T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}}$
    - bootstrap (next lesson)
  - Bernoulli data $F = Ber(\mu)$
    - confidence interval for proportions: $T = \frac{\bar{X}_n - \mu}{\sqrt{\bar{X}_n(1 - \bar{X}_n)} / \sqrt{n}}$

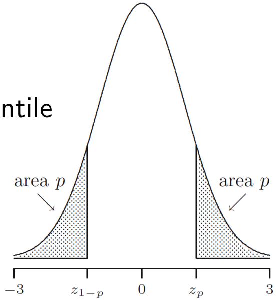
Critical values

Critical value

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

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

- The right tail is $P(Z \geq z_p) = 1 - P(Z \leq z_p) = 1 - \Phi(z_p)$
  - This is why Table B.1 of the textbook is given for $1 - \Phi()$
- $1 - \Phi(z_p) = p$ means $\Phi(z_p) = 1 - p$, i.e., z_p is the $(1-p)$th quantile
- By symmetry, $P(Z \geq z_p) = P(Z \leq -z_p) = p$, and then
  $$z_{1-p} = -z_p$$
  - E.g., $z_{0.975} = -z_{0.025} = -1.96$ and $z_{0.025} = -z_{.975} = 1.96$



CI for the mean: normal data with known variance

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

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

- 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 μ

One-sided confidence intervals

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

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

  Then (l_n, ∞) 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 μ

See R script

CI for the mean: normal data with unknown variance

- Use the unbiased estimator of σ^2 and its estimate:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

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

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

> 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
>
> $$f(x) = k_m\left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad \text{for } -\infty < x < \infty,$$
>
> where $k_m = \Gamma\left(\frac{m+1}{2}\right) / \left(\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}\right)$. This distribution is denoted by $t(m)$ and is referred to as the t-distribution with m *degrees of freedom*.

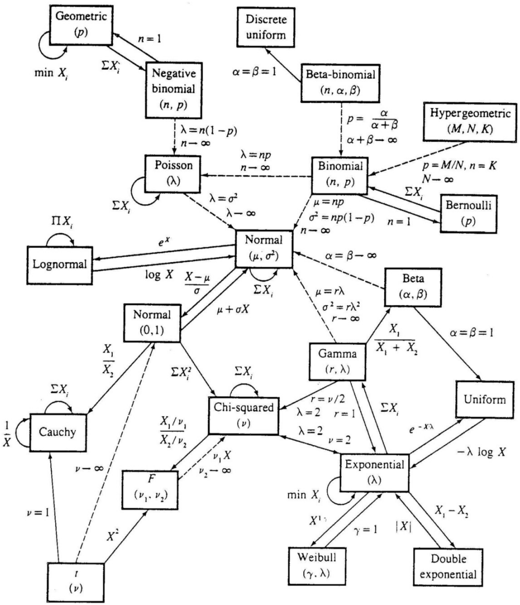
  - Student/Gosset t-distribution $X \sim t(m)$:
    - $E[X] = 0$ for $m \geq 2$, and $Var(X) = m/(m-2)$ for $m \geq 3$
    - For $m \to \infty$, $X \to \mathcal{N}(0,1)$

Some history on its discovery

See R script

Common distributions

- **Probability distributions at Wikipedia**
- **Probability distributions in R**
- C. Forbes, M. Evans,
  N. Hastings, B. Peacock (2010)
  Statistical Distributions, 4th Edition
  Wiley



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

CI for the mean: normal data with unknown variance

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

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 as z_p
- From the studentized mean:

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

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 μ

See R script

CI for the mean: general data with unknown variance

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

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

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

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

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 μ

See R script

Determining the sample size

- For a fixed α, 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 σ^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:

$$n \geq \left(2z_{\alpha/2}\frac{\sigma}{w}\right)^2$$

General form of Wald confidence intervals

$$\theta \in \hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}) \qquad \text{or} \qquad \theta \in \hat{\theta} \pm t_{\alpha/2} se(\hat{\theta})$$

- 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
- $B = \sum_{i=1}^n X_i \sim Bin(n, p)$ and $b = \sum_{i=1}^n x_i$ (number of observed successes)
  - For small n, build exact bounds (l_B, l_U) such that: [**Exact or Clopper–Pearson interval**]

$$l_B = \min_\theta \left\{ \sum_{x=B}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} \geq \alpha/2 \right\} \qquad u_B = \max_\theta \left\{ \sum_{x=0}^{B} \binom{n}{x} \theta^x (1-\theta)^{n-x} \geq \alpha/2 \right\}$$

- □ l_B is the smallest θ for which right tail $P(B \leq X) \geq {}^{\alpha}/_{2}$ for $X \sim Bin(n, \theta)$ *[left critical value]*
- □ u_B is the greatest θ for which left tail $P(X \leq B) \geq {}^{\alpha}/_{2}$ for $X \sim Bin(n, \theta)$ *[right critical value]*

$$P(l_B \leq p \leq u_B) = 1 - \alpha = \gamma$$

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

See R script

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) accuracy of classifier $y^+_\theta()$
- $B = \sum_{i=1}^n X_i \sim Bin(n,p)$ and $\bar{X}_n = B/n$
  - For large n, $Bin(n,p) \approx \mathcal{N}(np, np(1-p))$ for $0 \ll p \ll 1$ *[**De Moivre–Laplace**]*
    - and then $\bar{X}_n = B/n \approx \mathcal{N}(p, p(1-p)/n)$
    - $se(\bar{X}_n) = \sqrt{np(1-p)}/n \approx \sqrt{\bar{X}_n(1-\bar{X}_n)/n}$, because we don't known p
    - Consider $T = (\bar{X}_n - p)/se(\bar{X}_n) \approx \mathcal{N}(0,1)$ and then $P(-z_{\alpha/2} \leq T \leq z_{\alpha/2}) = \gamma$ implies:

$$P(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq p \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}) = 1 - \alpha = \gamma$$

      $(\bar{x}_n - z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}})$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for p
    - This is a Wald confidence interval!
  - Drawbacks: symmetric, large sample, skewness, etc. *[see **Wilson score interval** and others]*

See R script

Confidence intervals for simple linear regression coefficients

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

- We have $\hat{\beta} \sim \mathcal{N}(\beta, Var(\hat{\beta}))$ where $Var(\hat{\beta}) = \sigma^2/SXX$ is unknown *[see Lesson 20]*
- The Wald statistics is $t(n-2)$-distributed: *[proof omitted]*

$$\frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} \sim t(n-2)$$

- For $\gamma = 0.95$:

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

  and then a 95% confidence interval is: $\hat{\beta} \pm t_{n-2,0.025} se(\hat{\beta})$ where $se(\hat{\beta}) = \hat{\sigma}/\sqrt{SXX}$
- Similarly, we get for α, $\hat{\alpha} \pm t_{n-2,0.025} se(\hat{\alpha})$

See R script

Confidence intervals of fitted values

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

- For the fitted values $\hat{y} = \hat{\alpha} + \hat{\beta} x_0$ at x_0, a 95% confidence interval is:

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

where $se(\hat{y}) = \hat{\sigma}\sqrt{(\frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX})}$ *[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 $[\hat{y} + t_{n-2,0.025} se(\hat{y}), \hat{y} - t_{n-2,0.025} se(\hat{y})]$

See R script

Prediction intervals of fitted values

Simple linear regression: $Y_i = \alpha + \beta x_i + U_i$ with $\underline{U_i \sim \mathcal{N}(0, \sigma^2)}$ 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 *[See s4dsln.pdf Section 3.2]*

$$Var(\hat{V}) = \sigma^2 (1 + \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX})$$

- A 95% confidence interval is:

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

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

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

See R script

Optional reference

- 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.*