STAT 135, Concepts of Statistics

Helmut Pitters

Confidence intervals

Department of Statistics
University of California, Berkeley

February 16, 2017

Confidence intervals.

Example: Opinion polls. Consider town of $N = 25,000$ eligible voters. Taking a simple random sample $X_1, \ldots, X_{1600}$ of size $n = 1,600 = 40^2$ we find that 917 support Democrats. Suppose we try to estimate percentage

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = p$$

of people who support Democrats, where

$$x_i = \begin{cases} 1 & \text{if } i\text{th person votes for Democrats} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{2}{N}\mu\sum_{i=1}^{N} x_i + \mu^2 = p(1-p).$$

Confidence intervals.

Example: Opinion polls, cont. Idea: use

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{917}{1600} \approx 0.57$$

as estimator for p. We know (simple random sampling):

$$\mathbb{E}\hat{p} = p, \qquad \sigma_{\hat{p}} = \sqrt{\mathrm{Var}(\hat{p})} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \approx \frac{\sigma}{40},$$

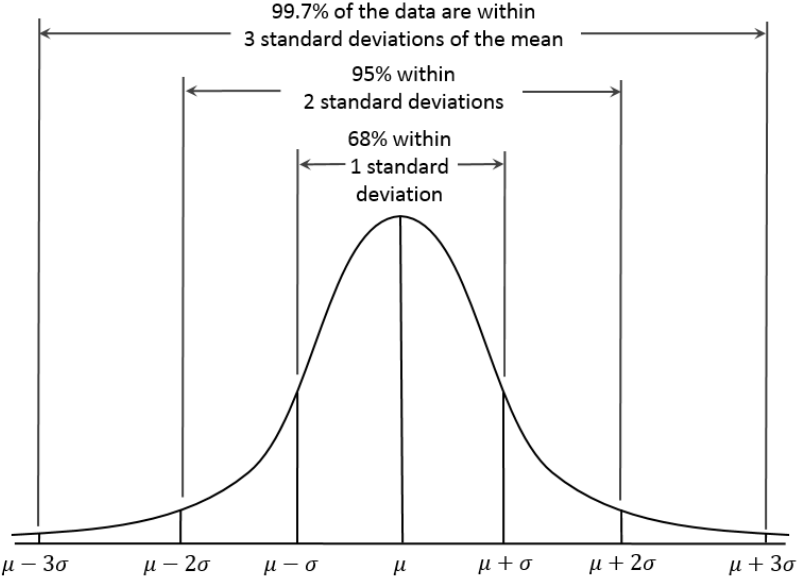
since sampling fraction $n/N = 1,600/25,000 = 0.064$ is small. Moreover, since $\sigma = \sqrt{p(1-p)}$ not known, estimate it by

$$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})} \approx 0.5,$$

i.e. estimate standard error $\sigma_{\hat{p}}$ by $0.0125 = 1.25\%$. Put differently, the percentage $\hat{p} \approx 0.57$ of Democrats in the sample is likely to be off the percentage of Democrats among all $25,000$ eligible voters by 1.25 percentage points or so.

Confidence intervals.

Recall the empirical rule for the standard Normal distribution.



Confidence intervals.

Example: Opinion polls, cont. Approximate $\hat{p}$ $(= \bar{X})$ by $\mathcal{N}(\hat{p}, \hat{p}(1-\hat{p})/n) = \mathcal{N}(0.57, 0.25/1600)$ (due to CLT). Hence

Table: Confidence intervals

$\hat{p} \pm 1.25\% = [0.55, 0.58]$	68.3%	confidence interval for p
$\hat{p} \pm 2 \times 1.25\% = [0.54, 0.6]$	95.5%	"
$\hat{p} \pm 3 \times 1.25\% = [0.53, 0.61]$	99.7%	"

Statistic.

Call a map $T(x_1, \ldots, x_n)$ of given data $x_1, \ldots, x_n$ a *statistic*.
Usually, regard these data as observed values of some random variables $X_1, \ldots, X_n$.
Moreover, in the case at hand, one needs to specify the (joint) distribution of the X_i that depends on some parameter, generically called $\theta > 0$.

Confidence intervals.

Consider a parameter θ that we want to estimate.[1] Suppose $a(X), b(X)$ are statistics such that

$$a(x) \leq b(x)$$

for all observations x generated by some random variable X, and that on seeing data $X = x$ we infer

$$a(X) \leq \theta \leq b(X).$$

If

$$\mathbb{P}\left\{a(X) \leq \theta \leq b(X)\right\} = 1 - \alpha$$

does not depend on θ, the random interval

$$[a(X), b(X)]$$

is called a $100(1-\alpha)\%$ *confidence interval*[2] for θ.

[1] E.g. think of θ as a population parameter in simple random sampling.

[2] Typically α is taken to be 0.05 or 0.01 so that probability that confidence interval contains θ is high.

Confidence intervals.

Example. $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with *unknown* μ and *known* σ^2.

Want to find 95% confidence interval for μ. Recall that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and suppose $a \leq b$ are such that

$$\mathbb{P}\left\{a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right\} = 1 - \alpha$$

which is equivalent to

$$\mathbb{P}\left\{\bar{X} - b\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - a\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

Due to symmetry of Normal distribution, length of confidence interval minimized for $-a = b$. Since $\Phi(1.96) = 0.975$,

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

t distribution

Definition

Let N, X be independent random variables such that

$$N \sim \mathcal{N}(0,1) \qquad X \sim \chi^2_n.$$

The distribution of

$$\frac{N}{\sqrt{X/n}}$$

is called a t *distribution with* n *degrees of freedom.*

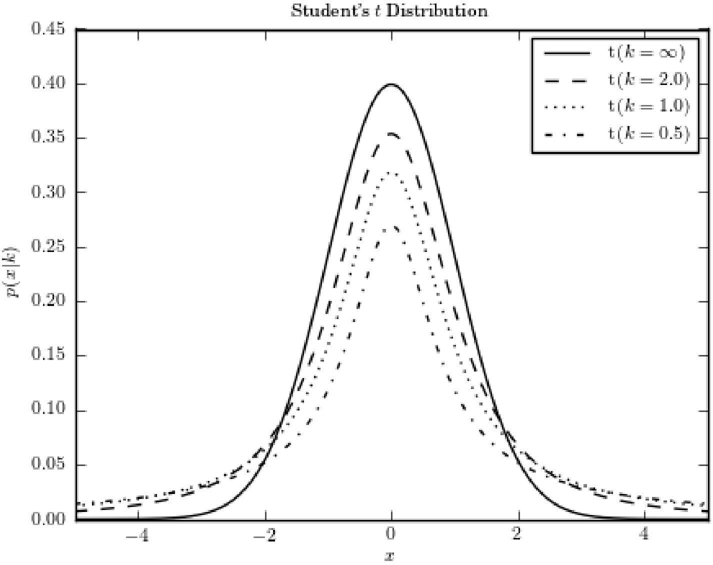
Fact

One can show that the t distribution has density

$$f(t) = \begin{cases} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} & t \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

Notice that the density of the t distribution is symmetric about 0.

t distribution



Confidence intervals.

Example. $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with *both* μ and σ^2 *unkown*. Want to find (shortest) 95% confidence interval for μ. From our results on distributions derived from Normal, we obtain:[3]

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n), \qquad (n-1)S^2/\sigma \sim \chi^2_{n-1},$$

and

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

follows Student's t-distribution with $n-1$ degrees of freedom.[4]

---

[3]Sample variance was defined to be $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

[4]However, if sample size n is large, a Normal approximation might be considered where σ^2 is estimated by the sample variance.

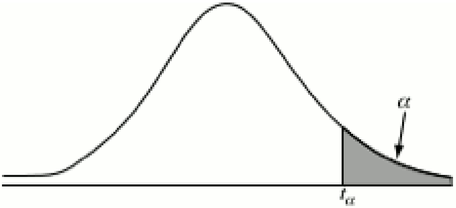
Confidence intervals.

Want 95% confidence interval which is symmetric about 0, i.e. want b such that

$$\mathbb{P}\{-b \leq T_{n-1} \leq b\} = 95\%,$$

where T_{n-1} follows Student's t distribution with $n-1$ degrees of freedom. Find from tabulated values of percentiles of t distribution...

Confidence intervals.



Values of α for one-tailed test and $\alpha/2$ for two-tailed test						
df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930

Confidence intervals.

Find from tabulated values of percentiles of t distribution **for** $n = 11$

$$\mathbb{P}\left\{-2.201 \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq 2.201\right\} = 95\%,$$

i.e.

$$\left[\bar{X} - 2.201\frac{S}{\sqrt{n}}, \bar{X} + 2.201\frac{S}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Confidence intervals.

How can we find a confidence interval for σ^2?

Recall

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Now define x_α by

$$\mathbb{P}\left\{X \leq x_\alpha\right\} = \alpha,$$

where X is some r.v. with distribution χ^2_{n-1}. Then

$$\begin{aligned} \alpha &= \mathbb{P}\left\{x_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq x_{1-\alpha/2}\right\} \\ &= \mathbb{P}\left\{\frac{(n-1)S^2}{x_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{x_{\alpha/2}}\right\}, \end{aligned}$$

thus

$$\left[\frac{(n-1)S^2}{x_{1-\alpha/2}}, \frac{(n-1)S^2}{x_{\alpha/2}}\right]$$

is a α-confidence interval for σ^2.