Confidence intervals (Section 9.2)

Set up: Same as before, i.e.,

Motivation: Estimator $\hat{\theta}$ is a single number that gives a plausible value of the unknown θ. But rarely the two will be equal. So, often it is preferable to give an interval of plausible values — a **confidence interval** (CI), which contains the unknown θ with a specified high probability.

Definition: An interval $[L, U]$ is a $100(1-\alpha)\%$ CI for θ if $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ are such that

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

- L and U are *random*, so the CI is *random*.
- Parameter θ is not random — it is unknown but fixed.
- $(1-\alpha)$ = *confidence coefficient* or *confidence level*.
- In practice, $(1-\alpha) = 0.90$ or 0.95 (most common) or 0.99.

A general method for constructing CI for θ

Step 1: Find an estimator $\hat{\theta}$ of θ that has a normal distribution with *known* variance, i.e., $\hat{\theta} \sim N(\theta, \mathrm{var}(\hat{\theta}))$.

Step 2: Standardize $\hat{\theta}$ to get Z, where

Step 3: Find a *critical point* $z_{\alpha/2}$ such that
$1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$

Thus, the $100(1-\alpha)\%$ CI is:

A general method, continued

Note 1: If the distribution of $\hat{\theta}$ is approximately normal, then the CI is also approximate.

Note 2: In case of MLE, if n is large, then often $\hat{\theta}$ approximately follows a $N(\theta, \hat{I}^{-1})$ distribution. In this case, an approximate $100(1-\alpha)\%$ CI for θ is:

Note 3: If the distribution is not normal or n is not large, we can use the method of bootstrap to construct a CI (later).

Confidence interval for population mean μ

Recall:

Case 1: The sample comes from a normal distribution with known variance. In this case,

Case 2: The sample comes from a any distribution, but n is large. In this case,

Ex: Suppose that an observed sample of size 20 from a $N(\mu, 10)$ population gives $\overline{x} = 2.45$ Find the 95% CI for μ.

Notice that this interval is *fixed* — it's a numerical interval. There is nothing random about it.

Q: Can we say that this observed interval contains the true value of μ with 95% probability?

So, how do we really interpret a CI?

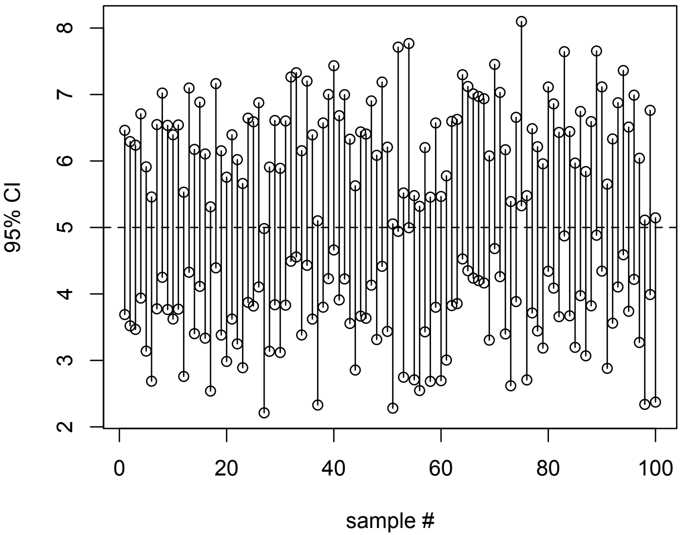
Interpretation of a CI

Recall:

- Usual long-term proportion interpretation of probability — i.e., if we repeat a large number of times the process of taking a random sample of size n from $N(\mu, \sigma^2)$ population and construct the CI using the above formula, then roughly 95% of times the observed CIs will be correct, i.e., it will capture the true value of μ.
- This CI formula gives an incorrect interval 5% (small) of the times.
- It is **wrong** to say that the observed interval contains the true value of μ with 95% probability. The CI either contains the true value of μ or it does not — we don't know what the case is.
- Thus, in a sense, we have 95% confidence in the CI formula — it gives the correct answer 95% of the times.

Lets use simulation to verify this interpretation.

- Draw a random sample of size 20 from a $N(5, 10)$ distribution and use the observed sample to construct a 95% CI for μ using the above formula.
- Repeat this procedure 10,000 times. The figure on the next page plots the constructed CIs for the first 100 samples.
- Find the proportion of times the CI captures the true value.


95% CI
2
3
4
5
6
7
8
0
20
40
60
80
100
sample #

```
# A function to simulate data from a N(mu, sigma^2)
# distribution and computing CI

conf.int <- function(mu, sigma, n, alpha){
x <- rnorm(n, mu, sigma)
ci <- mean(x) + c(-1,1) * qnorm(1-(alpha/2)) *
sigma/sqrt(n)
return(ci)
}

# Get one CI

mu <- 5
sigma <- sqrt(10)
n <- 20
alpha <- 0.05
```

```
# > conf.int(mu, sigma, n, alpha)
# [1] 3.520961 6.292768

# Repeat the process nsim times

nsim <- 10000
ci.mat <- replicate(nsim, conf.int(mu, sigma, n, alpha))

# > dim(ci.mat)
# [1]     2 10000

# The first 5 intervals

# > ci.mat[, 1:5]
         # [,1]     [,2]     [,3]     [,4]     [,5]
# [1,] 3.689654 3.519999 3.466402 3.937424 3.140117
# [2,] 6.461462 6.291807 6.238210 6.709231 5.911925
# >
# Graphing the first 100 intervals

plot(1:100, ci.mat[1, 1:100],
ylim=c(min(ci.mat[,1:100]), max(ci.mat[,1:100])),
xlab="sample #", ylab="95% CI", type="p")
points(1:100, ci.mat[2, 1:100])
for (i in 1:100) {
segments(i, ci.mat[1, i], i, ci.mat[2,i], lty=1)
}
abline(h=5, lty=2)

# Proportion of times the interval is correct

# > mean( (mu >= ci.mat[1,])*(mu <= ci.mat[2,]) )
# [1] 0.9502
# >
```

Confidence interval for a normal mean (known variance, cont'd)

Q: Given a random sample, which CI for μ would you prefer — a 95% CI or a 99% CI? (Note: `qnorm(0.975) = 1.959964`, `qnorm(0.995) = 2.575829`.)

Note the tradeoff:

- The **precision** of a CI is given by its **width**. The **accuracy** of a CI is given its **confidence level**.
- Higher confidence = lower precision (wider).
- The width of a 100% CI is: It is a useless interval — extremely "accurate" but extremely imprecise!

Q: What can we do to get a narrower CI without lowering the confidence?

- Width =
- Increase n to make CI more precise.

Choosing the sample size n:

- Let w = desired CI width for $1 - \alpha$ confidence.
- Margin of error = $w/2$
- Set the CI width to the desired width and solve for n to get

Ex: Suppose that we wish to estimate the mean CPU service time of a job and we wish to assert with 99% confidence that the estimated value is within less than 0.5 sec of the true value. Suppose that the past experience suggests that CPU service time is normally distributed with standard deviation $\sigma = 1.5$ sec. How many observations should we take?

Confidence interval for a normal mean (unknown variance)

- Unknown variance σ^2 is more realistic.
- Estimate σ^2 by sample variance, $S^2 =$

Pivot:

Result: $T \sim t_{n-1}$, i.e., a t-distribution $(n-1)$ degrees of freedom, instead of the $N(0,1)$ distribution.

- A t_{n-1}-distribution looks like a $N(0,1)$ but it has heavier tails. A heavier tail accounts for the fact that there is there is more uncertainty in T when S is used in place of σ
- When n is large, a t_{n-1}-distribution $\approx N(0,1)$.

Result: CI for μ: $\overline{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$

Proof:

- The t critical points are tabulated in the t-table. Alternatively, we can use `qt` function in `R`.
- Sample size calculation now becomes complicated than before because S needs to be known before data are collected.
- One option is to make an intelligent guess about S and be conservative (guess a larger value of S so that n larger than necessary is chosen).

Ex: If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? One way to do this is to compare mean time between keystrokes of the user trying to log in with that of the account owner. The intrusion is detected if there is a noticeable difference. The following times between keystrokes (in seconds) were recorded when a user typed the username and password:

$$0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34, 0.42, 0.09, 0.18, 0.46, 0.21$$

Find a 95% CI for mean time between keystrokes for the user trying to log in. Assume a normal distribution for the times.

```
x <- c(0.46, 0.38, 0.31, 0.24, 0.20, 0.31, 0.34,
0.42, 0.09, 0.18, 0.46, 0.21)
#> mean(x)
#[1] 0.3
#> sd(x)
#[1] 0.1183216
#>qt(0.975, 11)
#[1] 2.200985
```

Large sample CI for mean μ

Recall: When n is large, an approximate $100(1-\alpha)\%$ CI for mean μ of any population is

Ex: We wish to estimate the mean execution time of a program. The program was run 35 times on randomly selected inputs, and the sample mean and the sample standard deviation of the execution times were evaluated as 230 ms and 14 ms, respectively. Find a 95% CI for the true mean execution time μ.

Large sample CI for success proportion p

Population: $X \sim \text{Bernoulli }(p)$, where $p =$ proportion of successes in population; $p = E(X)$.

Sample data: $X_1, \ldots, X_n$. (Note: they are 0s and 1s).

Recall: Estimator for $p = \hat{p} =$ proportion of successes in the sample.
Also: Estimated var(X) = estimate of $p(1-p) =$

Result: An approximate CI for p: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Proof:

Ex: From a large population of RAM chips, a random sample of 50 is taken and a test carried out on each to see whether they perform correctly. In the test, only 20 chips are found to perform correctly. Find a 95% CI for p, the true proportion of chips that perform correctly.

Choosing the sample size n

- Width of CI =
- Let w = desired CI width for $1 - \alpha$ confidence.
- Margin of error = $w/2$
- Set CI width = desired width and solve for n to get

- This formula involves $\hat{p}$, which is not known before the experiment.
- One alternative: take $\hat{p}$= 0.5 because $\hat{p}(1 - \hat{p})$ is maximum when $\hat{p} = 0.5$. This strategy will yield a conservative values of n. (The sample size will be larger than necessary.)

Ex: Suppose we are planning a survey to estimate the proportion of American who approve of President Trump's job. We would like our estimate to be within 3% of the true proportion with 95% confidence. How much sample size should we take?

Two-sample problems

So far: One-sample problems — inference on parameter(s) of a single population.

Now: Two-sample problems — inference on parameters involving two populations.

Independent or paired samples?

Design 1: (Two independent samples)

Design 2: (Paired samples)

CI for $\mu_X - \mu_Y$ with paired design

- Data:
- Parameter of interest:
- Define: $D = X - Y$, $D_i = X_i - Y_i$, $i = 1, \dots, n$
-
- Apply one-sample procedure to the differences.

$100(1-\alpha)\%$ **CI for μ_D assuming $D_1, \dots, D_n \sim N(\mu_D, \sigma_D^2)$:**

Approximate $100(1-\alpha)\%$ CI for μ_D if n is large:

Q: What is the pivot here?

CI for $\mu_X - \mu_Y$ with two independent samples

Setup:

Scenario 1: No assumption regarding σ_X^2 and σ_Y^2 — they may be equal or unequal.

Estimators of parameters:

- μ_X:
- μ_Y:
- σ_X^2:
- σ_Y^2:
- $\mu_X - \mu_Y$:

Pivot:

- **Satterthwaite's approximation:** The distribution of the pivot can be approximated by t_ν distribution where

$$\nu = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{S_X^4}{n_X^2(n_X-1)} + \frac{S_Y^4}{n_Y^2(n_Y-1)}}$$

- Approximate $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$:

- When n_X, n_Y are large, the CI is:
- Is the normality assumption needed when n_X, n_Y are large?

Scenario 2: Assume common variance, i.e., $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Estimators of $(\mu_X, \mu_Y, \mu_X - \mu_Y)$:
- Estimate σ^2 using the *pooled* sample variance S_p^2:
- Is S_p^2 unbiased for σ^2?
- $SE(\overline{X} - \overline{Y}) =$
- Estimated SE:

Pivot:

- $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$:
- When n_X, n_Y are large, the CI is:
- Is the normality assumption needed when n_X, n_Y are large?

Large-sample CI for $p_X - p_Y$ with two independent samples

Setup:

- Estimators of $(p_X, p_Y, p_X - p_Y)$:
- $SE(\hat{p}_X - \hat{p}_Y) =$
- Estimated SE:
- **Pivot:**
- Approximate $100(1-\alpha)\%$ CI:

CI for a function of θ

Issue: Suppose (L, U) is a $100(1-\alpha)\%$ CI for θ. How to get a $100(1-\alpha)\%$ CI for $g(\theta)$, where g is a monotonically increasing function of θ?

Example 1

The data below show the sugar content (as a % of weight) of several national brands of children's and adults' cereals.

Children's cereals: 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.5

Adults' cereals: 20, 30.2, 2.2, 7.5, 4.4, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4

(a) Is it reasonable to assume that each sample comes from a normal distribution?

(b) Can the variances of the two distributions be assumed to be equal? Justify your answer.

(c) Compute an appropriate 95% confidence interval for difference in mean sugar contents of the two cereal types. What assumptions did you make, if any, to construct the CI?

(d) What do you conclude on the basis of your answer in (c)? Can we say that children's cereals have more sugar on average than adult cereals? If yes, by how much? Justify your answers.

Example 2

A study shows that 61 of 414 adults who grew up in a single-parent household report that they suffered at least one incident of abuse during childhood. By contrast, 74 of 501 adults who grew up in two-parent households report abuse.

(a) Is there a difference in single-parent and two-parent households when it comes to reporting abuse?

(b) What assumptions, if any, did you make to compute the interval in (a)? Do the assumptions seem reasonable?

Example 3

Consider the dataset stored in the file bp.txt. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods — a finger method and an arm method — from the same 200 patients.

(a) Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer.

(b) Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

(c) Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold?