Intervals

Lecture #3

Adi Sarid
Tel-Aviv University
updated: 2019-11-08

Reminder from previous lecture

Last lesson we talked about:

- Biased/unbiased estimators, variance of estimators (and standard errors)
- We discussed three methods of estimation
  1. Maximum Likelihood Estimation (MLE) $L(\theta) = \prod_i f(x_i; \theta)$; examples: Poisson, Normal.
  2. Bayesian method $\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$; examples: discrete/uniform for estimating a population proportion
  3. The moment method (use $EX^k = \sum x_i^k$ across $k = 1, 2, \ldots$)
- The central limit theorem: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to a normal disribution
- We started talking about statistical intervals $P(\hat{\Theta}_l < \theta < \hat{\Theta}_u) = 1 - \alpha$

Before we dive into statistical intervals

Two questions:

What is a good election survey?

What is a bad election survey?

We'll get back into this later on today.

Statistical Intervals (Montgomery chapter 8)

We discussed point estimates, however

- Even if everything works "properly" (a random sample, unbiased estimator), it is unlikely that we will reach the exact parameter value
- As the sample increases accuracy improves; but
- Sometimes we are interested in a *Confidence Interval*
- An interval of the form $\hat{\Theta}_l < \theta < \hat{\Theta}_u$ where
- The lower and upper bounds $\hat{\Theta}_l, \hat{\Theta}_u$ depend on the statistic $\hat{\Theta}$

In a probabalistic notation, we are looking for $\hat{\Theta}_l, \hat{\Theta}_u$ such that:

$$P(\hat{\Theta}_l < \theta < \hat{\Theta}_u) = 1 - \alpha$$

For $\alpha \in (0, 1)$. For example, when we set $\alpha = 0.05$, we call this a 95% confidence interval for θ.

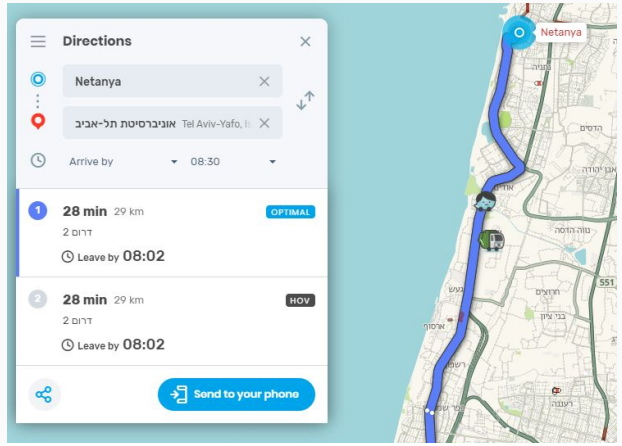
Motivation (example) road trippin' (1/3)

- Let's say we're planning a logistic operation
- We need to be in a specific place at a specific time
- We must not be late, but we can be a little early
- When should we depart?

Motivation (example) road trippin' (2/3)

Waze is cool, but… https://www.waze.com/livemap

- Not very robust for advance planning
- Specifically, we're only seeing a point estimate (average arrival time?) and not the distribution
- It's not that accurate either (30min to TLV in the rush hour?)



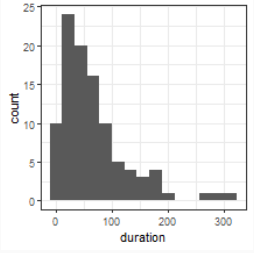
Motivation (example) road trippin' (3/3)

Assume we have Waze's raw data (needs to be **focused on the relevant time**, unbiased sample). We can compute a confidence interval.

```
set.seed(0)
drive_time <- tibble(duration = rexp(100, rate = 1/65))
# the rate is 1/65 cars per min. It means that it takes 65 minutes to get through
```

```
t.test(drive_time$duration,
       alternative = "two.sided",
       mu = mean(drive_time$duration))
```

```
##
##     One Sample t-test
##
## data:  drive_time$duration
## t = 0, df = 99, p-value = 1
## alternative hypothesis: true mean is not equal to 67.08955
## 95 percent confidence interval:
##  54.91621 79.26290
## sample estimates:
## mean of x
##  67.08955
```



To be 95% sure, we need to plan for **80 minutes' drive**.

Confidence Interval for Normal Distribution with Known Variance

We previously mentioned the central limit theorem and that

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Is normally distributed as $n \to \infty$. Hence:

$$P(z_{\alpha/2} < Z < x_{1-\alpha/2}) = 1 - \alpha$$

$$P(z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < z_{1-\alpha/2}) = 1 - \alpha$$

Using the fact that for the normal distribution $z_{1-\alpha/2} = -z_{\alpha/2}$:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

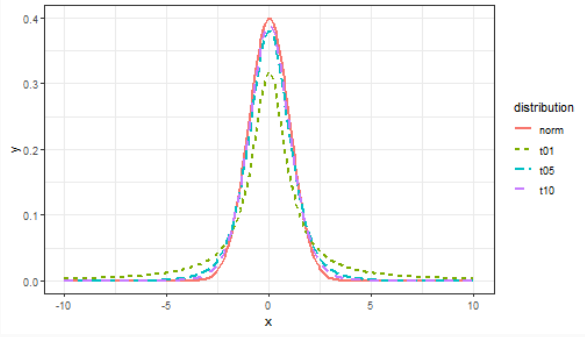
Confidence Interval for Normal Distribution with Unknown Variance

In this case, we use our estimator S to compute our statistic and confidence interval.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

The statistic T has a student's t-distribution with $n - 1$ degrees of freedom. I.e.:

$$P(-t_{\alpha/2,n} < T < t_{\alpha/2,n}) = 1 - \alpha$$



Back to the drive duration example

In the previous example we used `t.test`, let's break it down.

```r
n <- NROW(drive_time)

# t.test(drive_time$duration,
#        alternative = "two.sided",
#        mu = mean(drive_time$duration))

mean_duration <- mean(drive_time$duration)
sd_duration <- sd(drive_time$duration)
t_test_lims <- qt(p = c(0.025, 0.975), df = 99)

# This time, manually computed

mean_duration + t_test_lims*sd_duration/sqrt(100)
```

```
## [1] 54.91621 79.26290
```

Determining the sample size from a desired confidence range

If we want to have a confidence interval with a range not exceeding $\pm r$, we can use:

$$\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \leq 2r$$

Then,

$$\sqrt{n} \geq z_{\alpha/2}\frac{2\sigma}{2r}$$

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

Or (if variance is unknown):

$$n \geq \left(t_{\alpha/2,\text{n-1}}\frac{S}{r}\right)^2$$

What sample do we need in order to have a range r for the drive duration?

We want to have a 95% confidence interval in the drive duration, which is not longer than ± 4 minutes, i.e., $2r = 8$ minutes.

```r
desired_n <- ((qt(p = 0.975, df = 99))*sd_duration/ 4 )^2
desired_n
```

```
## [1] 926.1894
```

```r
set.seed(0) # illustration that this works
drive_time2 <- tibble(duration = rexp(desired_n, rate = 1/65))
t.test(drive_time2$duration,
       alternative = "two.sided",
       mu = mean(drive_time$duration))
```

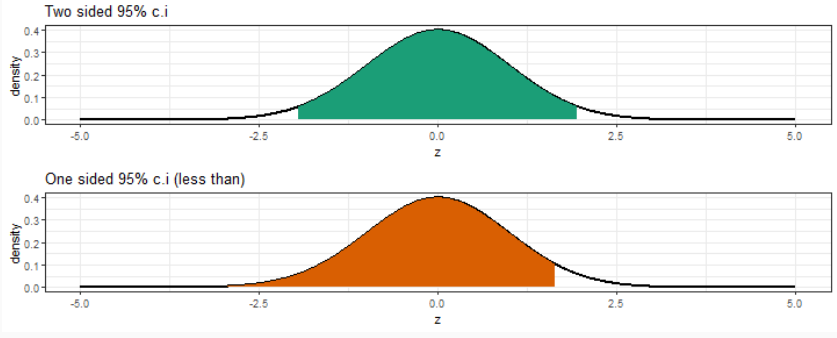
```
##
##     One Sample t-test
##
## data:  drive_time2$duration
## t = -0.29535, df = 925, p-value = 0.7678
## alternative hypothesis: true mean is not equal to 67.08955
## 95 percent confidence interval:
##  62.42485 70.53391
## sample estimates:
## mean of x
##  66.47938
```

One sided versus two sided confidence intervals

- We've discussed two-sided confidence intervals, i.e., $\theta \in [\hat{\Theta}_l, \hat{\Theta}_u]$
- Sometimes we prefer a one-sided confidence interval, for example when we one side is irrelevant, i.e. we want:
  - $P(\hat{\Theta}_l < \theta) = 1 - \alpha$, or
  - $P(\theta < \hat{\Theta}_u) = 1 - \alpha$
- This can be accomplished by using the right cutoff of the distribution, e.g.: z_α instead of using $z_{\alpha/2}$
  - $\bar{X} - z_\alpha \sigma / \sqrt{n} \leq \mu$, or
  - $\mu \leq \bar{X} + z_\alpha \sigma / \sqrt{n}$

One sided versus two sided confidence intervals - illlustration

```
normal_dist <- tibble(z = seq(-5,5,0.01)) %>%
   mutate(density = dnorm(z),
          cumulative = pnorm(z))
two_sided <- ggplot(normal_dist, aes(x = z, y = density)) +
   geom_line(size = 1) + geom_area(data = normal_dist %>%
                                     filter(cumulative <= 0.975 & cumulative >= 0.025),
                                   fill = "#1b9e77") +
   theme_bw() + ggtitle("Two sided 95% c.i")
one_sided_less <- ggplot(normal_dist, aes(x = z, y = density)) +
   geom_line(size = 1) + geom_area(data = normal_dist %>% filter(cumulative <= 0.95),
                                   fill = "#d95f02") +
   theme_bw() + ggtitle("One sided 95% c.i (less than)")
# one_sided_greater <- ggplot(normal_dist, aes(x = z, y = density)) +
#    geom_line(size = 1) + geom_area(data = normal_dist %>% filter(cumulative >= 0.05),
#                                    fill = "#7570b3") +
#    theme_bw() + ggtitle("One sided 95% c.i (greater than)")
gridExtra::grid.arrange(two_sided, one_sided_less, nrow = 2)
```



One sided versus two sided (in the example)

Discussion:

- In the example of drive duration what type of confidence interval would you use? why?
  - Two sided / One sided $\mu \leq C_L$ / One sided $\mu \geq C_U$

```r
t.test(drive_time$duration, alternative = "two.sided", mu = mean(drive_time$duration))
```

```
##
##     One Sample t-test
##
## data:  drive_time$duration
## t = 0, df = 99, p-value = 1
## alternative hypothesis: true mean is not equal to 67.08955
## 95 percent confidence interval:
##  54.91621 79.26290
## sample estimates:
## mean of x
##  67.08955
```

```r
t.test(drive_time$duration, alternative = "less", mu = mean(drive_time$duration))
```

```
##
##     One Sample t-test
##
## data:  drive_time$duration
## t = 0, df = 99, p-value = 0.5
## alternative hypothesis: true mean is less than 67.08955
## 95 percent confidence interval:
##     -Inf 77.2762
## sample estimates:
## mean of x
##  67.08955
```

General method to drive a confidence interval

We would like a general recipe that would work to generate confidence intervals for various types of distributions (not just the normal/student's t we've seen so far).

1. Find a statiatic $g(x_1, x_2, \ldots, x_n; \theta)$

2. The probability distribution of $g(x_1, x_2, \ldots, x_n; \theta)$ should not depend on θ (like in the Z case)

$$P(C_L \leq g(x_1, \ldots, x_n; \theta) \leq C_U) = 1 - \alpha$$

Since the probability does not depend on θ (property 2.), we can manipulate the expression inside the probability function:

$$P\Big(L(x_1, \ldots, x_n) \leq \theta \leq U(x_1, \ldots, x_n)\Big) = 1 - \alpha$$

Confidence intervals on variance and standard deviation of a normal

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $\mathrm{N}(\mu, \sigma)$, and set S^2 the sample variance $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ then

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

Has a chi-square χ^2 distribution with $n - 1$ degrees of freedom.

Alternatively, χ^2 can also be defined as a sum of squared **standard** normally distributed random variables $N_i \sim \mathrm{N}(0, 1)$ (the equivalence of these two definitions is out of our scope). Set

$$Y = N_1^2 + N_2^2 + \ldots + N_k^2$$

Then $Y \sim \chi_k^2$.

The mean and variance of a χ_k^2 distribution

What is the mean and variance of χ_k^2?

For this, I use the second definition:

$$Y = N_1^2 + N_2^2 + \ldots + N_k^2$$

Then, consider that $EN_i^2 = EN_i^2 - (EN_i)^2 + (EN_i)^2 = \sigma^2 + \mu^2 = 1 + 0$

- $EY = \sum EN_i^2 = k$

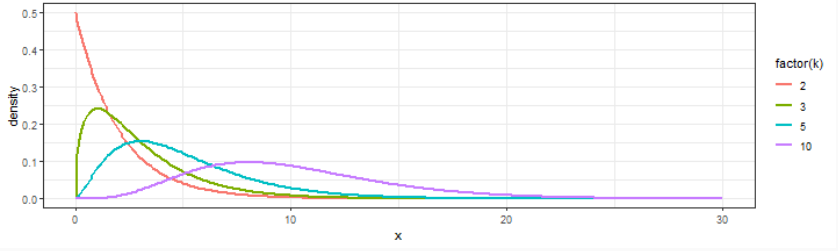
The fourth moment of a normal distribution can be computed directly, and is given by $3\sigma^4$, which for $\mathrm{N}(0, 1)$ is equal $3\sigma^4 = 3 \times 1^4 = 3$, hence

- $\mathrm{Var}(Y) = \sum \mathrm{Var}(N_i^2) = \sum EN_i^4 - (EN_i^2)^2 = \sum (3 - 1) = 2k$

Illustration of the χ^2_k for various k

```
chi_sq <- crossing(x = seq(0, 30, by = 0.1), k = c(2, 3, 5, 10)) %>%
   mutate(density = map2_dbl(x, k, dchisq))

ggplot(chi_sq, aes(x = x, y = density, color = factor(k))) +
   geom_line(size = 1) +
   theme_bw()
```



Question: What happens as $k \to \infty$ and why? (think about the central limit theorem)

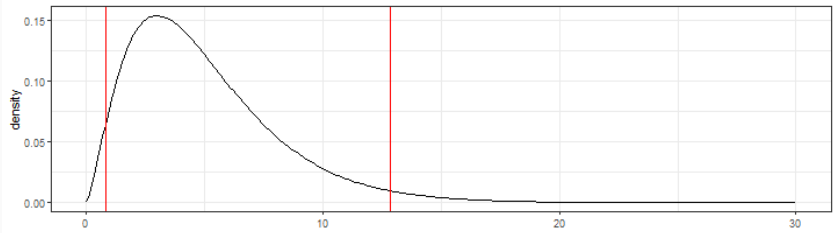
The χ^2 distribution is very useful in many statistical contexts. One of them is confidence intervals for σ^2.

A confidence interval for σ^2 and for σ

- If s^2 is the sample variance from a random sample of n observations
  - From a normal distribution with unknown variance σ^2
  - We can use the fact that $(n-1)s^2/\sigma^2$ is χ^2_{n-1} distributed for a confidence interval for σ^2 and for σ

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

```
df <- 5
chi_sq %>% filter(k == df) %>%
   ggplot(aes(x = x, y = density)) +
   geom_line() +
   geom_vline(xintercept = qchisq(c(0.025, 0.975), df), color = "red") +
   theme_bw()
```



Confidence interval for a population proportion

A common use of confidence intervals is for polling (survey results).

Who are you going to vote to in the next election?

- Let's say there is a candidate B.
- Survey results with $n = 500$ show that $\hat{p} = 200/500 = 40\%$.
- Would B cross the 50% threshold?

In essense we are dealing with a population proportion (the proportion of B's voters in the gen. pop.).

Consider the following random variable, using the central limit theorem (*show on whiteboard*) we can show it is normally distributed in the limit.

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

To show the CLT applies, we consider the fact that a Binomial random variable is a sum of Bernullis.

Would B be prime minister?

We can use a one-sided 95% confidence interval $\alpha = 0.05$ to see if B surpasses the 50%.

$$P\left(Z \geq z_\alpha\right) = P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq z_\alpha\right) = P\left(p \leq \hat{p} + z_{1-\alpha}\sqrt{p(1-p)/n}\right)$$

We replace $p(1-p)$ with $\hat{p}(1-\hat{p})$, similarly to how we replaced σ with s. If n is large enough, this yields a good approximation.

Our confidence interval is then:

$$p \leq \hat{p} + z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n} \implies p \leq 0.4 + 1.645 \times \sqrt{(0.4 \times 0.6)/500} \approx 43.7\%$$

Margin of error versus sample size

In the previous slide, if we were to produce a two-sided confidence interval, the result would have changed to the following range:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

The expression $\hat{p}(1-\hat{p})$ has it's maximum in 0.5, hence we can "enlarge" the range to:

$$\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{n}}$$

Setting $\alpha = 0.05$, $n = 500$, the term $\pm\frac{1.96}{2\sqrt{500}} \approx \pm 4.4\%$, which is what is commonly reported in surveys as an error up to 4.4%.

Examples here

Question 1: Why are there errors in election surveys if the margin of error is up to $\pm 4.4\%$?

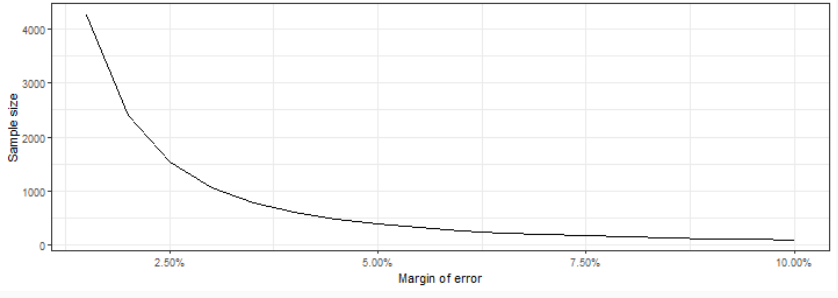
Question 2: What is the sampling method? (random? layers? "convinience" sampling?)

The required sample size as a function of the margin of error

We can plot the required sample size n as a function of the margin of error.

```
moe_n <- tibble(moe = seq(0.015, 0.1, by = 0.005),
                sample_size = (qnorm(0.975)/(2*moe))^2)

ggplot(moe_n, aes(moe, sample_size)) +
   geom_line() + scale_x_continuous(labels = scales::percent) +
   theme_bw() + xlab("Margin of error") + ylab("Sample size")
```



Prediction interval (1/3)

So far we've discussed confidence intervals, however, sometimes we are interested in a *prediction* interval, for a new observation.

- We have a sample of $x_1, \ldots, x_n$, random sample from a normal distribution
- We wish to predict the value x_{n+1} for a future observation
- The most obvious choice for a *point prediction* of x_{n+1} is $\bar{X}$
- The prediction error is given by $x_{n+1} - \bar{X}$ (unbiased prediction)
- The variance of the prediction error is $\mathrm{Var}(x_{n+1} - \bar{X}) = \sigma^2 + \sigma^2/n = \sigma^2(1 + 1/n)$
- Since $\bar{X} \sim \mathrm{N}(\mu, \sigma/\sqrt{n})$ and $x_{n+1} \sim \mathrm{N}(\mu, \sigma)$ and the two are independent, we have:
- $x_{n+1} - \bar{X} \sim \mathrm{N}(0, \sigma\sqrt{1 + 1/n})$

Prediction interval (2/3)

Now, we can follow the same steps we used for a confidence interval, replacing σ with s

$$T = \frac{x_{n+1} - \bar{X}}{s\sqrt{1 + \frac{1}{n}}}$$

Using the student's t distribution we can provide the following prediction interval

$$\bar{X} - t_{\alpha/2,n-1} \times s\sqrt{1 + \frac{1}{n}} \leq x_{n+1} \leq \bar{X} + t_{\alpha/2,n-1} \times s\sqrt{1 + \frac{1}{n}}$$

Prediction interval (3/3)

Important distinctions between confidence intervals and prediction intervals:

- In confidence intervals we are providing an interval for a **population parameter**
- In prediction intervals we are providing an interval for the **next actual value**
- The length of the confidence interval converges to 0
- The length of prediction interval converges to $2z_{\alpha/2}\sigma$.
- There will always be uncertainty associated with the next value, x_{n+1}, even when the average $\bar{X}$ is based on a very large sample, and is extremely close to μ.

Question: Reflecting back on the problem we started the lecture with (the drive duration problem). Should we have used a confidence interval or a prediction interval instead?