Point Estimation Methods and Intervals

Lecture #2

Adi Sarid
Tel-Aviv University
updated: 2019-11-03

Reminder from previous lecture (1/3)

Last lesson we talked about:

- How data analysis is conducted (import -> tidy -> transform, visualize, model -> communicate)
- About R, very broadly (RStudio, Rmd, scripts, functions, packages)
- What a tidy data set looks like (rows = observations, columns = variables, aka features)
- Variable types (numeric, date, logical, factor - category, ordinal)
- Demonstration of the grammer of graphics (`ggplot2`)

Then, we also discussed **theory**:

- What are point estimates, i.e.:
  - θ is a *population parameter*
  - Estimated by the statistic $\hat{\Theta}$, a *point estimator*
  - When it is computed (from a sample), it is called a *point estimate*

Reminder from previous lecture (2/3)

- Desired properties of point estimates are:
  - Unbiased: $E\hat{\Theta} - \theta = 0$
  - Low variance: $V(\hat{\Theta})$ as low as possible

I've shown that:

- The average $\bar{X} = \sum_{i=1}^{n} x_i$ is unbiased:
  - $E[\bar{X}] = \mu$
- Its variance is $V(\bar{X}) = \frac{\sigma^2}{n}$
- In fact, it is the *Minimum Variance Unbiased Estimate* (proof out of scope)
- We've also seen the bias-variance decomposition, i.e. the Mean Square Error is decomposed into:

$$E[(\hat{\Theta} - \theta)^2] = V(\hat{\Theta}) + E(\hat{\Theta} - \theta)^2$$

Reminder from previous lecture (3/3)

We talked about *quantiles*/*percentiles*/*median*:

- A quantile $q(f)$ of a sample is the **value** for which a specified fraction f of the data values is $\leq q(f)$
- Note that quartile, quantile, percentile are related/interchangable:
  - 2 quartile = .5 quantile = 50 percentile = median

To sum up

- The average: $\bar{X} = \sum_{i=1}^{n} x_i$ is an **unbiased** estimator of $E[X] = \mu$
- Sample variance: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$ **unbiased** estimator of $V(X) = \sigma^2$
  - $V(X) = E(X - E[X])^2 = EX^2 - (EX)^2$
- Standard deviation: S is a **biased** estimator for σ (can't enjoy both worlds)

Why S^2 is an unbiased estimator to σ^2?

S^2 is unbiased (you actually saw this in the exercise):

$$ES^2 = \frac{1}{n-1} E \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} E \sum_{i=1}^{n} (x_i^2 + \bar{X}^2 - 2\bar{X}x_i)$$

$$= \frac{1}{n-1} E \left(\sum_{i=1}^{n} x_i^2 - n\bar{X}^2 \right) = \frac{1}{n-1} \left[\sum_{i=1}^{n} Ex_i^2 - nE(\bar{X}^2) \right]$$

Using the fact that $E(x_i^2) = \mu^2 + \sigma^2$ and that $E(\bar{X}^2) = \mu^2 + \sigma^2/n$ we have:

$$E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) = \sigma^2$$

$$\square$$

Why is S a biased estimator to σ?

We need to show that $ES \neq \sigma$. Let

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

Consider that $0 < \mathrm{Var}(S) = E[S^2] - (E[S])^2$ (this is true for any RV, specifically S in this case)

Hence

$$(ES)^2 < E[S^2] \Leftrightarrow ES < \sqrt{E[S^2]} = \sigma$$

$$\square$$

In certain cases, we can directly compute this bias (i.e. what is $ES - \sigma$), for example, if we assume $X \sim N(\mu, \sigma)$, then:

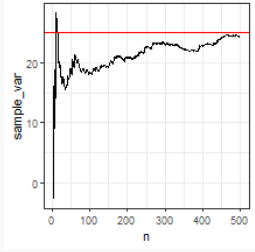
$$\sigma - E(S) \cong \frac{\sigma}{4n}$$

(see here).

Estimating S^2 - demonstration

Take a normal distribution with

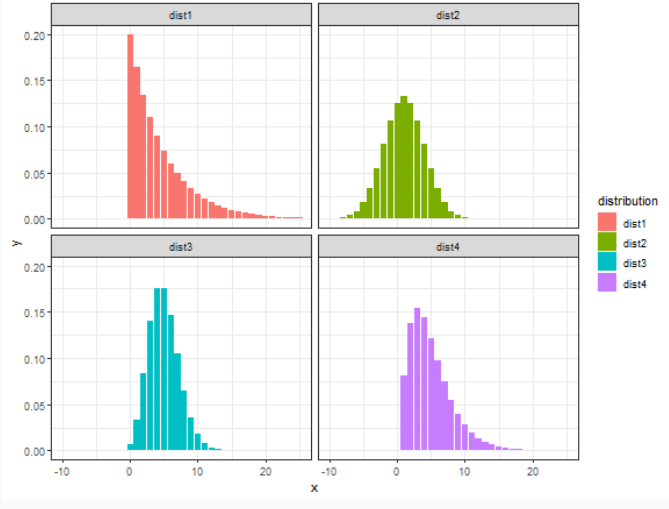
```
set.seed(0)
rv <- tibble(x = rnorm(500, mean = 0, sd = 5)) %>%
  mutate(cumsum_x = cumsum(x),
         n = seq_along(x),
         average_x = cumsum_x/n,
         sample_var = (1/(n-1))*cumsum(x^2-average_x^2)) %>%
  slice(-1)

ggplot(rv, aes(x = n, y = sample_var)) +
  geom_line() +
  geom_hline(aes(yintercept=25), color = "red") +
  theme_bw()
```



Short quiz - identify the distributions

Before we continue, **identify the distributions using the plot**, and write them down on a piece of paper, in pairs.

Here are your options: Normal, Chi-square, Bernulli, Exponential, Poisson.



The Central Limit Theorem

If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

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

As $n \to \infty$ is the standard normal distribution $N(0, 1)$

For most purpuses, $n \geq 30$ is considered "large enough" (as a rule-of-thumb).

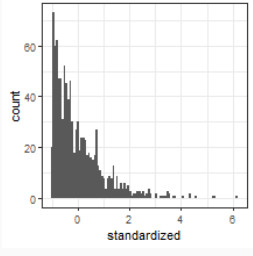
Lets look at an example of exponential distribution: $\text{Exp}(\lambda = 1)$. **What are μ and σ?**

The distribution of $\bar{X}$ with varying n

With $\text{Exp}(\lambda = 1)$ we have $\mu = \sigma = 1/\lambda = 1$

Try increasing `sample_size` and see what happens to the chart.

```
sample_size <- 1
lambda <- 1

suppressMessages(
rv_binom <- matrix(rexp(n = sample_size*1000, rate = lambda), nrow = sample_size, ncol = 1000) %>%
  as_tibble(.name_repair = "unique") %>%
  pivot_longer(cols = everything(), names_to = "var", values_to = "value") %>%
  group_by(var) %>%
  summarize(mean = mean(value)) %>%
  mutate(standardized = (mean - 1/lambda)/((1/lambda)/sqrt(sample_size))))

ggplot(rv_binom, aes(x = standardized)) +
  geom_histogram(bins = 100) +
  theme_bw()
```



Methods of point estimation

So far we discussed some estimators (i.e., $\bar{X}$, S^2), and the desired properties they hold. But how can we find additional estimators (new statistics)?

We will show three methods:

- Maximum Liklihood Estimation (MLE)
- The Bayesian method
- Moment method

Maximum Likelihood Estimation (MLE)

An important and very common approach to solving estimation problems in statistics. The idea is as follows:

If you want to estimate some population parameter θ, use the most **likely** value.

Let $f(x;\theta)$ represent the density function of X. Given a sample $x_1, \ldots, x_n$ in which x_i's are independent, we can write the liklihood of the sample:

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \times \ldots \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

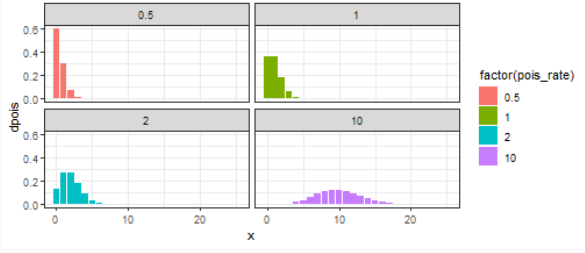
- The product (multiplication) can be used thanks to the assumption of independence of the sampled observations.
- $L(\theta)$ is a function of θ (only) once the sample has been set.
- In the discrete case we will use $P(X_1 = x_1, \ldots, X_n = x_n; \theta)$. The likelihood is the same as the probability of obtaining the sample.

MLE example - Poisson distribution (discrete)

The Poisson distribution is used to represent a counting processes. I.e., the number of accumulated events is distributed Poisson.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

```
poiss_demonstration <- crossing(x = 0:25, pois_rate = c(0.5, 1, 2, 10)) %>%
   mutate(dpois = map2_dbl(x, pois_rate, dpois))

ggplot(poiss_demonstration, aes(x = x, y = dpois, fill = factor(pois_rate))) +
   geom_col() +
   facet_wrap(~pois_rate) +
   theme_bw()
```



MLE example - Poisson distribution (discrete) - cont.

Assume we sample the number of arrivals to the hospital during n weekdays: $x_1, \ldots, x_n$. The number of arrivals is distributed with a $\text{Poisson}(\lambda)$ distribution.

Then, the likelihood is:

$$L(\lambda) = \prod_{i=1}^{n} (\lambda^{x_i} e^{-\lambda}) / x_i!$$

Taking logarithm we have:

$$\log L(\lambda) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda \log(e) - \log(x_i!))$$

Now we require an extremum, i.e. $d \log L(\lambda) / d\lambda = 0$:

$$\sum_{i=1}^{n} x_i / \lambda^* - n = 0 \Longrightarrow \lambda^* = \bar{X}$$

This is indeed a maximum $\frac{d^2 \log L(\lambda)}{d\lambda^2} = -\left(\sum_{i=1}^{n} \frac{x_i}{\lambda}\right) < 0$

MLE example - Bernulli distribution

Let's assume we conducted n experiments each with probability p for success and $q = 1 - p$ for failure.
These are Bernulli i.i.d variables.

$$B_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1 - p \end{cases}$$

Assume we got v success, $n - v$ failures.

- What is the Liklihood?
- What is the log-Liklihood?
- Find the optimal $\hat{p}$.

You will do this in the homework.

Maximum Liklihood Estimate - Normal distribution

The previous examples dealt with discrete distributions. What happens in the continuous case?

Assume $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ i.i.d distributed. We sample $x_1, \ldots, x_n$ and we are looking for an MLE for μ and σ^2:

The likelihood:

$$L(\mu) = \prod_{i=1}^{n} \left(\frac{e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

Taking log we get:

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

$$\frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} \implies \mu^* = \bar{X}$$

Maximum Liklihood Estimate - Normal distribution - cont.

Now, derivative by σ we obtain:

$$\frac{\partial \log L(\mu, \sigma^2)}{\partial(\sigma^2)} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

From here we get:

$$(\sigma^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

As you have seen, $(\sigma^*)^2$ is a **biased** estimator to the variance (the unbiased estimator s^2 had $n-1$ in the denominator).

The Bayesian method (reminder: Bayes' rule)

First, a reminder about **Bayes' rule**.

The definition of conditional probability is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

For $P(B) > 0$. Then, since: $P(A \cap B) = P(B \cap A)$ we also have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

If we have a partition $E_1, \ldots, E_k$ (i.e., $\cup E_i = \Omega$, $E_i \cap E_j = \phi$) we get:

$$P(B) = P(B|E_1)P(E_1) + \ldots + P(B|E_k)P(E_k)$$

And finally, we arrive at the general form of Bayes' theorem:

$$P(E_1|B) = \frac{P(B|E_1)P(E_1)}{P(B|E_1)P(E_1) + \ldots + P(B|E_k)P(E_k)}$$

Bayesian Methods of Estimation

We have a parameter θ for a population with a distribution $f(x|\theta)$. We assume some prior distribution $\pi(\theta)$ (represents our belief on the unknown value of θ).

Assume a sample $x = (x_1, \ldots, x_n)$, then the sampling distribution is $f(x|\theta)$.

We want to use the sampling distribution (obtained from data), and our prior belief in order to yield a *posterior* distribution, i.e., an estimate, of θ.

We use Bayes' rule

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)}$$

Where:

$$g(x) = \begin{cases} \sum_\theta f(x|\theta)\pi(\theta) & \theta \text{ discrete} \\ \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta & \theta \text{ continuous} \end{cases}$$

Bayesian Methods of Estimation - Example (1/3)

We have a machine with either 10% or 20% defects (w.p. 0.6 or 0.4 respectively).

p	**0.1**	**0.2**
$\pi(p)$	0.6	0.4

We sampled the assembly line with two items, $n = 2$, and find out x defects $x \in \{0, 1, 2\}$. What is the Bayesian estimator $\pi(p|x)$ for p?

$$\pi(p|x) = \frac{f(x|p)\pi(p)}{g(x)}$$

Let's compute all the elements. x is binomially distributed. Its probability function is:

$$f(x|p) = \binom{2}{x} p^x (1-p)^{2-x}$$

The marginal distribution of x is:

$$g(x) = \sum_p \pi(p) \times f(x|p) = \pi(0.1) \times f(x|0.1) + \pi(0.2) \times f(x|0.2) = 0.6 \times f(x|0.1) + 0.4 \times f(x|0.2)$$

Bayesian Methods of Estimation - Example (2/3)

We can easily compute the binomial probability using `R`.

```
f_x_p <- function(x, p){
   dbinom(x, size = 2, prob = p)
}

pi_p <- function(p){
   if (p == 0.1){
      0.6
   } else if (p == 0.2){
      0.4
   } else {
      0
   }
}

g_x <- function(x){
 pi_p(0.1)*f_x_p(x, 0.1) +
      pi_p(0.2)*f_x_p(x, 0.2)
}

pi_p_x <- function(p, x){
   (f_x_p(x, p)*pi_p(p)) / (g_x(x))
}
```

Bayesian Method of Estimation - Example (3/3)

Now we can compute $\pi(p|x)$ for any possible combination of p and sample yielding x defects.

```
res <- tibble(p = c(0.1, 0.1, 0.1, 0.2, 0.2, 0.2),
              x = c(0:2, 0:2)) %>%
   mutate(pi_p_x = map2_dbl(p, x, pi_p_x))

knitr::kable(res, format = "html", padding = 3)
```

p	x	pi_p_x
0.1	0	0.6549865
0.1	1	0.4576271
0.1	2	0.2727273
0.2	0	0.3450135
0.2	1	0.5423729
0.2	2	0.7272727

Mean of the posterior distribution

The mean of the posterior distribution $p^* = E_p[\pi(p|x)]$ is called the **Bayes estimate of** p. In the previous example, if we find one defect (1 of 2 sampled) we get:

$$p^* = (0.1) \times 0.46 + (0.2) \times 0.54 \approx 0.154$$

Note the big difference from the classical estimator (the average). The average in this case is $\bar{p} = x/n = 1/2 = 0.5$.

Bayesian methods - continuous example

This time, we are going to assume that p can obtain any value within $[0, 1]$.

First, answer (write on a piece of paper) in pairs:

- What distribution should we use a-priori to describe the value of p?

  1. Normal with $\mu = 0.5$ and $\sigma = 0.001$
  2. Uniform $[0, 1]$
  3. Exponential $\mathrm{Exp}(\lambda = 0.001)$

- What is $\pi(p)$?

  - According to your answer to the previous question write $\pi(p) = \ldots$?

Bayesian methods - continuous example - cont.

Uniform distribution! The others have a positive probability for values higher than 1, and since our parameter p reflects distribution, it can only obtain values in $[0, 1]$.

$$\pi(p) = \begin{cases} 1 & p \in [0, 1] \\ 0 & \text{Otherwise} \end{cases}$$

$$\pi(p|x) = \frac{f(x|p)\pi(p)}{\int f(x|p)\pi(p)dp} = \frac{\binom{2}{x} p^x (1-p)^{2-x}}{\int_0^1 \binom{2}{x} p^x (1-p)^{2-x} dp}$$

Take for example $x = 1$ then:

$$p^* = \int_0^1 \pi(p|x=1)dp = 3\binom{2}{1}\int_0^1 p \times p^1 (1-p)^{2-1} dp = 1/2$$

- Exactly the same we would have got with $\bar{p} = x/n = 1/2$.

Method of Moments Estimation

In this method, we use knowledge about the moments of the distribution (i.e. $E[X], E[X^2], E[X^3], \ldots$), express the parameters as functions of these moments, and then use the sample moments to compute our estimator.

Let $f(x)$ be a density function for $X_1, \ldots, X_n$, then:

Moment	**Continuous**	**Discrete**	**Sample**
$E[X]$	$\int x f(x) dx$	$\sum_k kP(X=k)$	$(1/n)\sum_{i=1}^n x_i$
$E[X^2]$	$\int x^2 f(x) dx$	$\sum_k k^2 P(X=k)$	$(1/n)\sum_{i=1}^n x_i^2$
$E[X^m]$	$\int x^m f(x) dx$	$\sum_k k^m P(X=k)$	$(1/n)\sum_{i=1}^n x_i^m$

The sample mean $\bar{X} = (1/n)\sum_{i=1}^n x_i$ is the moment estimator of the population mean.

Method of Moments - Example - the Normal disribution

The first moment is $E[X] = \mu$, hence $\hat{\mu} = \bar{X}$ is the moment estimator for the population mean.

The second moment is $E[X^2] = \mu^2 + \sigma^2$ require:

$$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Set $\mu^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{\sum x_i^2 - (\sum x_i)^2/n}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

Again, this is a biased estimator for σ^2, usually we will use s^2 seen earlier.

In the homework, you will be requested to use the moment method on two distributions (the exponential distribution and another distribution) to find estimators for parameters of the distributions.

Statistical Intervals

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.

Sanity check

What would be a 100% confidence interval?

- I.e., what would be $\hat{\Theta}_l, \hat{\Theta}_u$ such that:

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

Setting $\hat{\Theta}_l = -\infty, \hat{\Theta}_u = \infty$ gives us a 100% confidence interval (i.e., the θ is a real number).

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:

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

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

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

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

Example - 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,

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

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

Ever wondered why surveys have $n = 500$ respondents? it comes from a $\pm 4.4\%$ margin of error with a 95% confidence interval (using a binomial distribution).

Maybe in the next homework.

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.:

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