Fundamentals of Mathematical Statistics and Matrix Algebra

Philip Leifeld

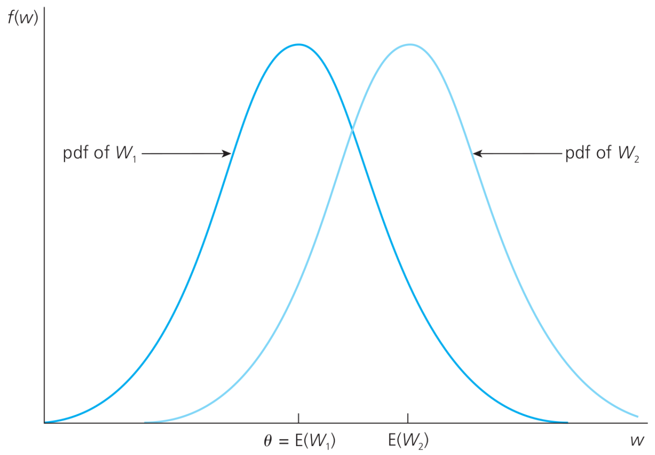
GV903: Advanced Research Methods, Week 4

University of Essex

1. Properties of Estimators

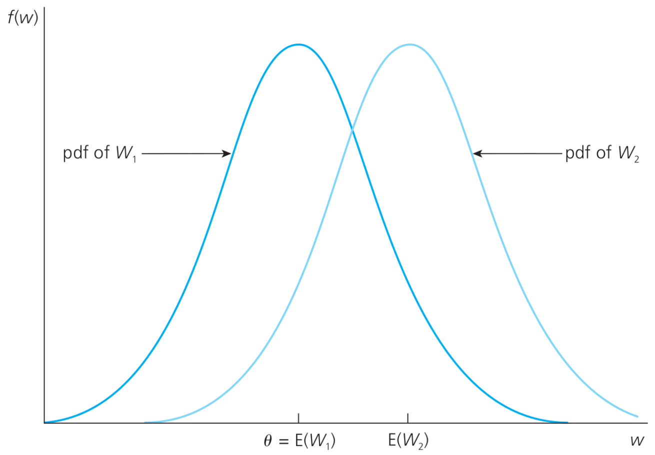
Bias of an Estimator

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



Bias of an Estimator

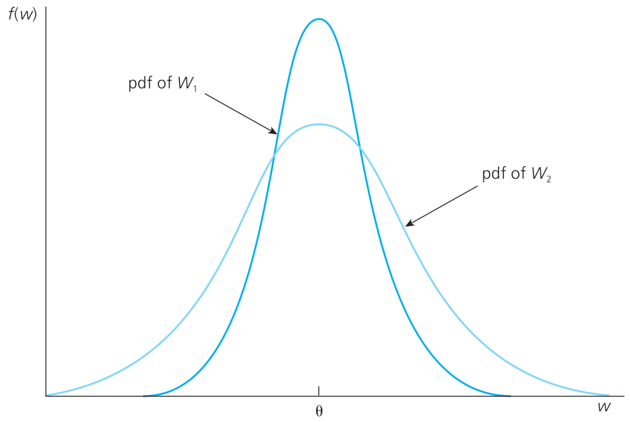
FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



Unbiasedness: $E(W) = \theta$. Bias$(W) = E(W) - \theta$.

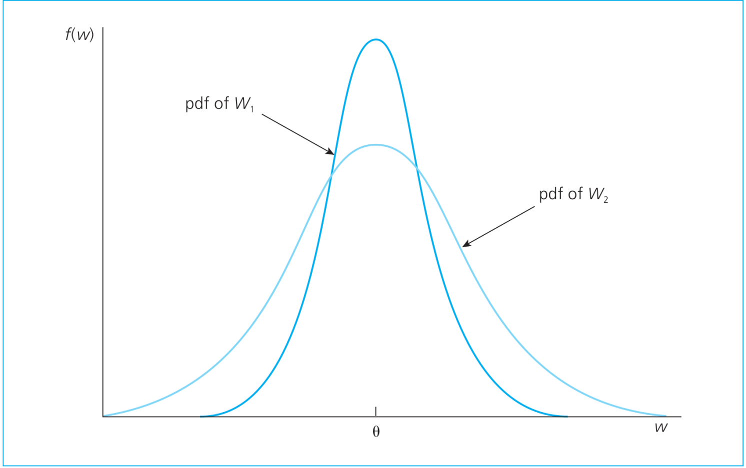
Efficiency of an Estimator

FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



Efficiency of an Estimator

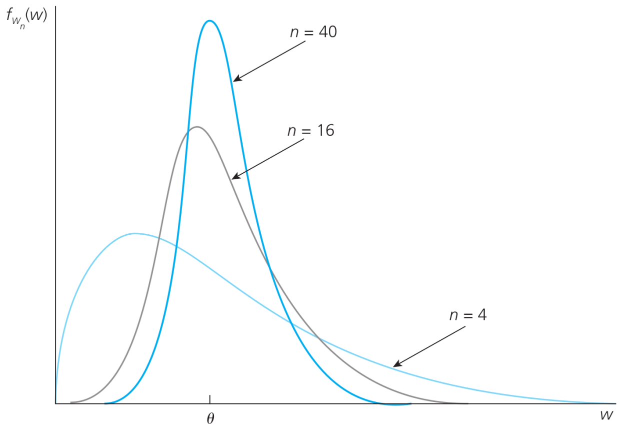
FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



We want unbiased *and* efficient estimators, but there is sometimes a trade-off.
Assess both with the *mean squared error*: $\text{MSE}(\hat{\theta}) = \text{Var}_\theta(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2$

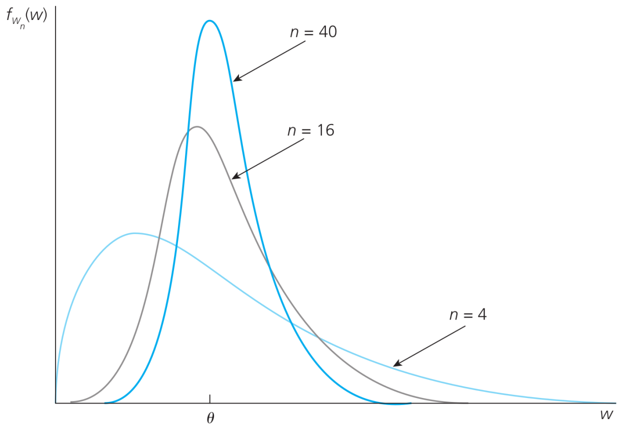
Consistency of an Estimator

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



Consistency of an Estimator

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



Law of large numbers: we get arbitrarily close to μ as the sample size increases.
$\text{plim}(W_n) = \theta$

2. Implementing Functions in R

Implementing Functions in R: The Mean

Computing the mean: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

In R:

```
numbers <- c(2, 6, 4, 7, 12)
mean(numbers)
```

[1] 6.2

Re-implementing the `mean` function:

```
mean2 <- function(v) {
  return(sum(v) / length(v))
}
mean2(numbers)
```

[1] 6.2

Implementing Functions in R: The Median

Median (for ordered x): $m(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & n \text{ even} \end{cases}$

Implementing Functions in R: The Median

Median (for ordered x): $m(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & n \text{ even} \end{cases}$

```r
median2 <- function(v) {
  v <- sort(v)
  if (length(v) %% 2 == 1) {  # uneven length
    v[(length(v) + 1) / 2]
  } else {                    # even length
    (v[length(v) / 2] + v[(length(v) / 2) + 1]) / 2
  }
}
median2(numbers)
```

[1] 6

```r
median2(c(4, 7, 10, 13))
```

[1] 8.5

Implementing Functions in R: Standard Deviation

Population SD: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$. Sample SD: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

Implementing Functions in R: Standard Deviation

Population SD: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$. Sample SD: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

```r
sd2 <- function(v, sample = TRUE) {
  differences <- numeric(length(v))
  for (i in 1:length(v)) {
    differences[i] <- (v[i] - mean(v))^2
  }
  if (sample == TRUE) {
    return(sqrt(sum(differences) / (length(v) - 1)))
  } else {
    return(sqrt(sum(differences) / length(v)))
  }
}
sd2(numbers, sample = TRUE)  # default of sd()
```

```
[1] 3.768289
```

```r
sd2(numbers, sample = FALSE)
```

```
[1] 3.37046
```

Exercise

The trace of a matrix is defined as: $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

The trace is only defined for quadratic matrices.

In R, the trace can be computed as `sum(diag(mat))`.

Can you write a function that returns the trace of a matrix without using the `sum` and `diag` functions? Use `for`-loops.

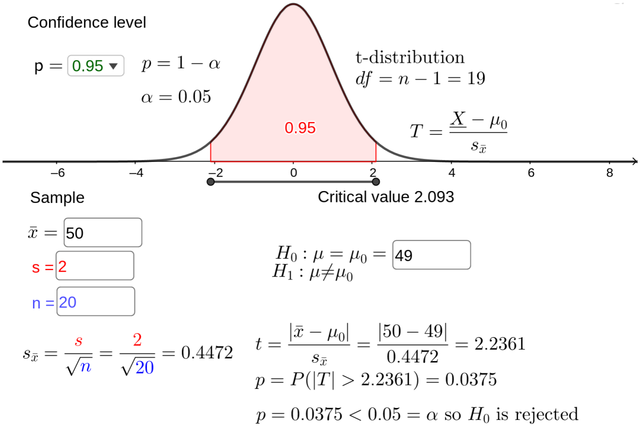
Solution

```r
trace <- function(mat) {
 if (nrow(mat) != ncol(mat)) {
   stop("Matrix is not quadratic!")
 }
 tr <- 0
 for (i in 1:nrow(mat)) {
   for (j in 1:ncol(mat)) {
     if (i == j) {
       tr <- tr + mat[i, j]
     }
   }
 }
 return(tr)
}
trace(matrix(1:36, ncol = 6))
## [1] 111
```

3. Hypothesis Testing

The Logic of Hypothesis Testing

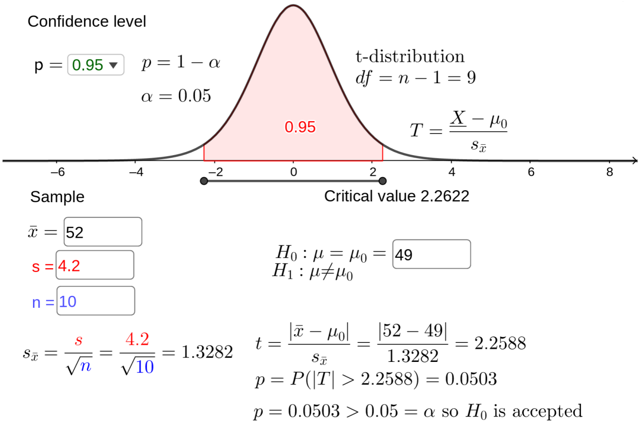
1. Define H_0 and H_1.
2. Select a useful distribution.
3. Select an α confidence level.
4. One- or two-sided test?
5. Critical value in distribution? $c = F^{-1}(\alpha_{\frac{1}{2}}; df)$
6. Compute test statistic based on sample: $t = \frac{\bar{x}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}}{\text{se}(\bar{y})}$
7. Reject H_0 if $|t| > c$
8. p value: Probability of obtaining test results at least as extreme as observed. $p = 2(1 - F(|t|))$



https://www.geogebra.org/m/fbq2xhrt

The Logic of Hypothesis Testing

1. Define H_0 and H_1.
2. Select a useful distribution.
3. Select an α confidence level.
4. One- or two-sided test?
5. Critical value in distribution?
   $c = F^{-1}(\alpha_{\frac{1}{2}}; df)$
6. Compute test statistic based on sample:
   $t = \frac{\bar{x}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}}{\mathrm{se}(\bar{y})}$
7. Reject H_0 if $|t| > c$
8. p value: Probability of obtaining test results at least as extreme as observed.
   $p = 2(1 - F(|t|))$



https://www.geogebra.org/m/fbq2xhrt

Let's Apply this Logic to the Poisson Distribution!

The average number of protests in a fictitious city per year is 35. After the outbreak of Covid-19, the number of protests in 2021 is 24. Is this evidence that Covid makes people protest less?

Let's Apply this Logic to the Poisson Distribution!

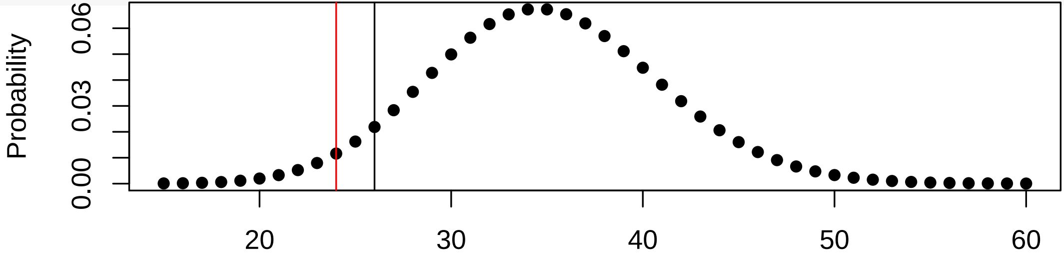
The average number of protests in a fictitious city per year is 35. After the outbreak of Covid-19, the number of protests in 2021 is 24. Is this evidence that Covid makes people protest less?

1. Protests $\sim$ Poisson(λ).
2. H_0: $\lambda = 35$. H_1: $\lambda < 35$.
3. $\alpha = 0.05$.
4. One-sided test (smaller than before!).
5. Critical value: $\mathrm{F}^{-1}_{\mathrm{Poisson}}(0.05, \lambda = 35) = 26$.
6. Test statistic: $t = 24 < c \Rightarrow$ reject H_0.
7. $p = P(Y \leq 24) = \mathrm{F}_{\mathrm{Poisson}}(t, \lambda = 35) = \sum_{y=0}^{24} \frac{35^y e^{-35}}{y!} = 0.0323741$.

Now in R. . .

```
x <- 15:60
y <- dpois(x, lambda = 35)
plot(x, y, pch = 16, xlab = "Count", ylab = "Probability")
abline(v = qpois(p = 1 - 0.95, lambda = 35))
abline(v = 24, col = "red")
ppois(q = 24, lambda = 35) # p-value
## [1] 0.03237411

sum(sapply(0:24, function(y) {(35^y * exp(-35)) / factorial(y)}))
## [1] 0.03237411
```



Implementing Functions in R: Two-sided t Test

Test statistic: $t = \frac{\bar{y}}{\frac{s}{\sqrt{n}}} = \frac{\bar{y}}{\text{se}(\bar{y})}$. Critical value: $c = F^{-1}(\alpha_{\frac{1}{2}}; df)$

p value: $1 - \Phi(t)$. Confidence interval: $[\mu - c \cdot se; \mu + c \cdot se]$

Implementing Functions in R: Two-sided t Test

Test statistic: $t = \frac{\bar{y}}{\frac{s}{\sqrt{n}}} = \frac{\bar{y}}{\mathrm{se}(\bar{y})}$. Critical value: $c = F^{-1}(\alpha_{\frac{1}{2}}; df)$

p value: $1 - \Phi(t)$. Confidence interval: $[\mu - c \cdot se; \mu + c \cdot se]$

```
t.test2 <- function(x, alpha = 0.05) {
  m <- mean(x)
  se <- sd(x) / sqrt(length(x))
  t <- m / se
  cval <- qt(1 - (alpha / 2), df = length(x) - 1)
  pval <- 2 * (1 - abs(pt(t, df = length(x) - 1)))
  cat("|t| > c: ", abs(t), " > ", cval, ", p = ",
      pval, ".\nEstimate: ", m, " [", m - cval * se,
      "; ", m + cval * se, "].", sep = "")
}
t.test2(c(2, 3, 3, 2, 1, -2, 1))
## |t| > c: 2.199707 > 2.446912, p = 0.07013051.
## Estimate: 1.428571 [-0.1605442; 3.017687].
```

Exercise

Investigative journalists have uncovered eight random party donations the *Party for the Elites* received in year t_2, from a larger pool of donation transactions. These donations have the following volume (in thousand £): 140, 190, 23, 5, 98, 55, 300, 221.

In year t_1, the party received an average of £114,000 from their donors. How confident can we be that the donations in year t_2 are due to a willingness to spend more than in the previous year (rather than random fluctuation)?

1. Conduct a hypothesis test manually with the appropriate distribution. What is the critical value? What is the test statistic? Use $\alpha = 0.05$. You can look up the critical value in R using the respective quantile distribution function, or use Appendix G in Wooldridge.
2. Compute the estimate and CI. Compute the p-value.
3. Repeat these steps in R to check if you found the right solution.

Solution

Subtract previous mean 114 from all values:
26, 76, 91, -109, -16, -59, 186, 107.

Mean changes: $\overline{\Delta y} = \frac{26+76-91-109-16-59+186+107}{8} = 15$

SD: $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-15)^2}{8-1}} = 103.2307$

$t = \frac{15}{\frac{103.2307}{\sqrt{8}}} = 0.4109864$

$c = F^{-1}(\alpha = 0.05; df = 8-1) = 1.894579$

$t \not> c \Rightarrow$ Not a significant increase in donations!

$p = P(c > 0.4109864|\mathrm{H}_0) = 1 - F_z(0.4109864) = 0.3405412$

$[\mu - c \cdot \frac{103.2307}{\sqrt{8}}; \mu + c \cdot \frac{103.2307}{\sqrt{8}}] = [-54.14752; 84.14752]$

Solution in R

```r
v <- c(140, 190, 23, 5, 98, 55, 300, 221)
v <- v - 114
m <- sum(v) / length(v)
l <- length(v)
s <- numeric(l)
for (i in 1:l) {
  s[i] <- (v[i] - m)^2
}
s <- sqrt(sum(s) / (l - 1))
s
## [1] 103.2307

tval <- m / (s / sqrt(8))
tval
## [1] 0.4109864
```

Solution in R

```r
cval <- qt(0.95, df = l - 1)
cval
## [1] 1.894579

1 - pnorm(tval) # p-value (using the standard normal distribution)
## [1] 0.3405412

1 - pt(tval, df = l - 1) # p-value (using the t distribution)
## [1] 0.3466863

m - (cval * (s / sqrt(8)))
## [1] -54.14748

m + (cval * (s / sqrt(8)))
## [1] 84.14748
```

Solution in R

Or simply using the `t.test` function:

```
t.test(v, alternative = "greater")
##
##  One Sample t-test
##
## data:  v
## t = 0.41099, df = 7, p-value = 0.3467
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -54.14748       Inf
## sample estimates:
## mean of x
##        15
```

Some Notes about the First Assignment

- Weeks 2–5 are covered.
- Several tasks with several sub-questions each.
- Some tasks will need to be solved manually with equations.
- Some tasks will need to be solved in R.
- Some questions will require writing a short text of about 100–200 words.
- I will give a few points for typesetting the answers in LaTeX, possibly with R code inserted using `knitr` in RStudio. Details will be included in the assignment.
- The `R` scripts discussed in the lab sessions are relevant.
- All readings up to (including) Week 5 are relevant, not just the lecture contents.
- I will *not* ask you to provide mathematical proofs this time.
- Tasks may be similar to the tasks from the lectures, perhaps a bit more complex.