Sufficiency and the Rao–Blackwell Theorem

Last modified: 2026-11-12 22:02:32 (PST)

A statistic summarizes a sample, and some summaries lose nothing that matters for a parameter. This page defines such summaries, called sufficient statistics, and shows how to use one to improve an estimator without making it more biased.

Sufficient statistics

Definition 0.1 (Statistic)**.** A **statistic** is a function of the sample, $T(\underset{\sim}{X})$, that does not depend on any unknown parameter. Its value at the observed data, $T(\underset{\sim}{x})$, is a number or vector we can compute.
Source: Wikipedia contributors [1].

Remark (Estimators are statistics)

Remark 0.1 (Estimators are statistics). Every estimator is a statistic, one chosen to land close to an estimand. This page asks which statistics lose no information about the parameter, and how to use one to improve an estimator.

Definition 0.2 (Sufficient statistic)**.** Let $\underset{\sim}{X}$ be a discrete sample whose distribution depends on a parameter θ. A statistic $T(\underset{\sim}{X})$ is **sufficient** for θ if, for every sample $\underset{\sim}{x}$ and every value t, the conditional probability of the sample given the statistic is the same for every value of θ at which the statistic takes the value t with positive probability:

$$\Pr_\theta(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) \text{ does not depend on } \theta \quad \text{wherever } \Pr_\theta(T(\underset{\sim}{X}) = t) > 0 \tag{1}$$

Source: Wikipedia contributors [2].

Remark (Source: Sufficient statistic)

Remark 0.2 (Source: Sufficient statistic). Fisher [3] introduced sufficiency; Casella and Berger [4], Def. 6.2.1, p. 272.

Remark (What sufficiency says)

Remark 0.3 (What sufficiency says). Once we know $T(\underset{\sim}{X})$, the remaining randomness in the sample has the same distribution whatever the parameter is, so the sample carries no further information about θ beyond $T(\underset{\sim}{X})$. A statistic can therefore stand in for the whole sample when we estimate θ.
The sample itself, taking $T(\underset{\sim}{X}) = \underset{\sim}{X}$, is always sufficient but does not reduce the data; a useful sufficient statistic is a much shorter summary, such as a sum or a mean.
Sufficiency by itself does not give a best estimator: it says which summaries lose nothing, not which function of them to use. The Rao–Blackwell theorem below shows how to improve a given unbiased estimator.
For a continuous sample, the definition, the factorization theorem and the conditional expectations below use densities and integrals instead of probability mass functions and sums.

Theorem 0.1 (Factorization theorem)**.** *Let $\mathbb{f}(\underset{\sim}{x};\theta)$ be the joint probability mass function of a discrete sample $\underset{\sim}{X}$. A statistic $T(\underset{\sim}{X})$ is sufficient for θ if and only if there are nonnegative functions g and h such that, for every sample $\underset{\sim}{x}$ and every θ,*

$$\mathbb{f}(\underset{\sim}{x};\theta) = g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x}) \tag{2}$$

where $h(\underset{\sim}{x})$ does not depend on θ.
Source: Wikipedia contributors [2].

Proof

Proof. Let $\underset{\sim}{x}$ be a value of the sample $\underset{\sim}{X}$ and write $t \overset{\text{def}}{=} T(\underset{\sim}{x})$. The event $\underset{\sim}{X} = \underset{\sim}{x}$ lies inside the event $T(\underset{\sim}{X}) = t$.
Sufficient implies a factorization. Define $g(t;\theta) \overset{\text{def}}{=} \Pr_\theta(T(\underset{\sim}{X}) = t)$ and let $h(\underset{\sim}{x})$ be the conditional probability $\Pr_\theta(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t)$ at any θ with $\Pr_\theta(T(\underset{\sim}{X}) = t) > 0$. By sufficiency, $h(\underset{\sim}{x})$ does not depend on which such θ we use. (If $\Pr_\theta(T(\underset{\sim}{X}) = t) = 0$ for every θ, set $h(\underset{\sim}{x}) \overset{\text{def}}{=} 0$.) Both functions are nonnegative, and at a θ with $\Pr_\theta(T(\underset{\sim}{X}) = t) = 0$ both sides of Equation 2 are zero, because the event $\underset{\sim}{X} = \underset{\sim}{x}$ lies inside the event $T(\underset{\sim}{X}) = t$. At every other θ,

$$
\begin{aligned}
\mathbb{f}(\underset{\sim}{x};\theta) &= \Pr_\theta(\underset{\sim}{X}=\underset{\sim}{x}) && \text{(definition of the joint mass function)}\\
&= \Pr_\theta(\underset{\sim}{X}=\underset{\sim}{x},\ T(\underset{\sim}{X})=t) && \text{(the first event is inside the second)}\\
&= \Pr_\theta(T(\underset{\sim}{X})=t)\ \Pr_\theta(\underset{\sim}{X}=\underset{\sim}{x}\mid T(\underset{\sim}{X})=t) && \text{(conditional probability)}\\
&= g(t;\theta)\,h(\underset{\sim}{x}) && \text{(definitions of } g \text{ and } h\text{)}
\end{aligned}
$$

A factorization implies sufficient. Suppose Equation 2 holds, and let t be a value with $\Pr_\theta(T(\underset{\sim}{X})=t)>0$. The event $T(\underset{\sim}{X})=t$ is the union of the disjoint events $\underset{\sim}{X}=\underset{\sim}{y}$ over the samples $\underset{\sim}{y}$ with $T(\underset{\sim}{y})=t$, so

$$
\begin{aligned}
\Pr_\theta(T(\underset{\sim}{X})=t) &= \sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} \mathbb{f}(\underset{\sim}{y};\theta) && \text{(disjoint events add)}\\
&= \sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} g(T(\underset{\sim}{y});\theta)\,h(\underset{\sim}{y}) && \text{(factorization)}\\
&= \sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} g(t;\theta)\,h(\underset{\sim}{y}) && \text{(each term has } T(\underset{\sim}{y})=t\text{)}\\
&= g(t;\theta) \sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} h(\underset{\sim}{y}) && \text{(factor out the common term)}
\end{aligned}
$$

This is positive, so $g(t;\theta)>0$. For a sample $\underset{\sim}{x}$ with $T(\underset{\sim}{x})=t$,

$$
\begin{aligned}
\Pr_\theta(\underset{\sim}{X}=\underset{\sim}{x}\mid T(\underset{\sim}{X})=t) &= \frac{\mathbb{f}(\underset{\sim}{x};\theta)}{\Pr_\theta(T(\underset{\sim}{X})=t)} && \text{(conditional probability)}\\
&= \frac{g(t;\theta)\,h(\underset{\sim}{x})}{g(t;\theta)\sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} h(\underset{\sim}{y})} && \text{(factorization, and the display above)}\\
&= \frac{h(\underset{\sim}{x})}{\sum_{\underset{\sim}{y}:\,T(\underset{\sim}{y})=t} h(\underset{\sim}{y})} && \text{(divide by } g(t;\theta)>0\text{)}
\end{aligned}
$$

which does not depend on θ. For a sample $\underset{\sim}{x}$ with $T(\underset{\sim}{x})\neq t$ the conditional probability is zero, which does not depend on θ either. So $T(\underset{\sim}{X})$ is sufficient. □

ℹ Remark (Source: Factorization theorem)

Remark 0.4 (Source: Factorization theorem). Halmos and Savage [5], whom Casella and Berger [4], Thm. 6.2.6, p. 276, and its footnote credit with the general form, and Neyman (1935) with an earlier one.

Example 0.1 (The sum is sufficient for a Poisson rate)**.** Let $X_1, \dots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$ and $S \stackrel{\text{def}}{=} \sum_{i=1}^n X_i$. The joint mass function of the sample is

$$\begin{aligned} \mathbb{f}(\underset{\sim}{x}; \lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} && \text{(independent Poisson mass functions)} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i} \prod_{i=1}^n \frac{1}{x_i!} && \text{(collect the factors)} \end{aligned}$$

This has the form of Equation 2 with $g(t;\lambda) = e^{-n\lambda}\lambda^t$ and $h(\underset{\sim}{x}) = \prod_{i=1}^n 1/x_i!$, where $t = \sum_{i=1}^n x_i$. So the sum S is sufficient for λ by Theorem 0.1.

The Rao–Blackwell theorem

Definition 0.1 (Rao–Blackwellized estimator)**.** Let $\hat{\theta}$ be an estimator of θ, and let $T(\underset{\sim}{X})$ be a sufficient statistic for θ. The **Rao–Blackwellized estimator** is the conditional expectation of $\hat{\theta}$ given the statistic:

$$\hat{\theta}_{\text{RB}} \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{\theta} \mid T(\underset{\sim}{X})\right] \tag{3}$$

Lemma 0.1 (The Rao–Blackwellized estimator is a statistic)**.** *Let $T(\underset{\sim}{X})$ be a sufficient statistic for θ and $\hat{\theta}$ an estimator of θ with finite expectation. Then $\mathbb{E}\left[\hat{\theta} \mid T(\underset{\sim}{X})\right]$ does not depend on θ, so it is a statistic and an estimator.*

Proof

Proof. Take a discrete sample, and choose a value t that the statistic $T(\underset{\sim}{X})$ takes with positive probability. Write $\hat{\theta} = \hat{\theta}(\underset{\sim}{X})$.

$$\mathbb{E}\left[\hat{\theta} \mid T(\underset{\sim}{X}) = t\right] = \sum_{\underset{\sim}{x}} \hat{\theta}(\underset{\sim}{x}) \Pr_\theta(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) \quad \text{(conditional expectation)}$$

Each term involves $\hat{\theta}(\underset{\sim}{x})$, which does not depend on θ, and the conditional probability, which does not depend on θ by Equation 1. So the sum does not depend on θ. □

Theorem 0.1 (Rao–Blackwellizing preserves unbiasedness)**.** *If $\hat{\theta}$ is an unbiased estimator of θ and $T(\underset{\sim}{X})$ is a sufficient statistic for θ, then $\mathbb{E}\left[\hat{\theta} \mid T(\underset{\sim}{X})\right]$ is also unbiased for θ.*

Proof

Proof. We use the law of iterated expectations:

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}\Big[\hat{\theta} \mid T(\underset{\sim}{X})\Big]\Big] &= \mathbb{E}\Big[\hat{\theta}\Big] && \text{(law of iterated expectations)}\\ &= \theta && (\hat{\theta} \text{ is unbiased})\end{aligned}$$

□

Theorem 0.2 (Rao–Blackwell theorem)**.** *Let $\hat{\theta}$ be an unbiased estimator of θ with finite variance, and let $T(\underset{\sim}{X})$ be a sufficient statistic for θ. Then the Rao–Blackwellized estimator is unbiased and has variance no larger than that of $\hat{\theta}$, for every θ:*

$$\operatorname{Var}\Big(\mathbb{E}\Big[\hat{\theta} \mid T(\underset{\sim}{X})\Big]\Big) \leq \operatorname{Var}\Big(\hat{\theta}\Big) \tag{4}$$

Source: [6], [7], [4], Thm. 7.3.17, p. 342, [8].

Proof

Proof. Unbiasedness is Theorem 0.1. For the variance, we use the law of total variance:

$$\begin{aligned}\operatorname{Var}\Big(\hat{\theta}\Big) &= \operatorname{Var}\Big(\mathbb{E}\Big[\hat{\theta} \mid T(\underset{\sim}{X})\Big]\Big) + \mathbb{E}\Big[\operatorname{Var}\Big(\hat{\theta} \mid T(\underset{\sim}{X})\Big)\Big] && \text{(law of total variance)}\\ &\geq \operatorname{Var}\Big(\mathbb{E}\Big[\hat{\theta} \mid T(\underset{\sim}{X})\Big]\Big) && \text{(a variance is nonnegative)}\end{aligned}$$

□

Exercise 0.1 (Conditional probability of a zero given the Poisson sum)**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \text{Pois}(\lambda)$ and $S \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i$. Find $\Pr(X_1 = 0 \mid S = t)$ for each $t = 0, 1, 2, \ldots$. You may use that $S \sim \text{Pois}(n\lambda)$, by the sum of independent Poisson variables and induction on n.

Solution

Solution 0.1. The event $X_1 = 0$ and $S = t$ happens exactly when $X_1 = 0$ and $X_2 + \cdots + X_n = t$, and $X_2 + \cdots + X_n \sim \text{Pois}((n-1)\lambda)$ is independent of X_1.

$$
\begin{aligned}
\Pr(X_1 = 0 \mid S = t) &= \frac{\Pr(X_1 = 0,\ S = t)}{\Pr(S = t)} && \text{(conditional probability)} \\
&= \frac{\Pr(X_1 = 0,\ X_2 + \cdots + X_n = t)}{\Pr(S = t)} && \text{(on } X_1 = 0,\ S \text{ is that sum)} \\
&= \frac{\Pr(X_1 = 0)\Pr(X_2 + \cdots + X_n = t)}{\Pr(S = t)} && \text{(independence)} \\
&= \frac{e^{-\lambda} \cdot e^{-(n-1)\lambda}((n-1)\lambda)^t/t!}{e^{-n\lambda}(n\lambda)^t/t!} && \text{(Poisson mass functions)} \\
&= \frac{e^{-n\lambda}((n-1)\lambda)^t/t!}{e^{-n\lambda}(n\lambda)^t/t!} && (e^{-\lambda}e^{-(n-1)\lambda} = e^{-n\lambda}) \\
&= \frac{((n-1)\lambda)^t}{(n\lambda)^t} && \text{(cancel } e^{-n\lambda}/t!) \\
&= \left(\frac{n-1}{n}\right)^t && \text{(cancel } \lambda^t, \text{ combine powers)}
\end{aligned}
$$

Lemma 0.2 (Conditional probability of a zero given the Poisson sum)**.** *For* $X_1, \ldots, X_n \ \sim_{\text{iid}}\ \text{Pois}(\lambda)$ *and* $S \overset{def}{=} \sum_{i=1}^n X_i$,

$$
\Pr(X_1 = 0 \mid S = t) = \left(\frac{n-1}{n}\right)^t \tag{5}
$$

for every $t = 0, 1, 2, \ldots$.

Proof

Proof. This is the solution to Exercise 0.1. □

Example 0.1 (Rao–Blackwellizing an estimate of the chance of a zero count)**.** The `birthwt` data record the number of previous premature labors (`ptl`) for 189 women in a study at Baystate Medical Center in Springfield, Massachusetts [9]; the data ship with the R package `MASS` (`MASS::birthwt`), and this site has a copy at `data/birthwt.csv`. Treat counts like these as a Poisson sample, $X_1, \ldots, X_n \ \sim_{\text{iid}}\ \text{Pois}(\lambda)$, with λ equal to the sample mean of `ptl`, 0.196.

The estimand is the probability of a zero count, $\pi = \Pr(X_i = 0) = e^{-\lambda}$. Index the model by π instead of λ (the map $\lambda \mapsto e^{-\lambda}$ is one-to-one); then the sum $S = \sum_{i=1}^n X_i$ is still sufficient, by Example 0.1.

The crude estimator is the indicator that the first observation is zero, $\hat{\pi}_1 = \mathbb{1}\{X_1 = 0\}$, which uses one observation and is unbiased for π. Its Rao–Blackwellization is

$$
\begin{aligned}
\hat{\pi}_2 &= \mathbb{E}[\hat{\pi}_1 \mid S] && \text{(definition of the Rao–Blackwellized estimator)} \\
&= \Pr(X_1 = 0 \mid S) && \text{(the mean of an indicator is a probability)} \\
&= \left(\frac{n-1}{n}\right)^S && \text{(conditional probability of a zero)}
\end{aligned}
$$

By Lemma 0.2, the conditional probability in the last line is the one in Equation 5. For $n = 20$ we draw 10,000 samples, compute both estimators on each, and report each one's sample mean (with a 95% confidence interval) and variance (with a 95% confidence interval from the normal approximation to the sampling distribution of a sample variance).

```r
birthwt <- read.csv("data/birthwt.csv")

# The Poisson rate: the observed mean number of previous premature labors.
lambda <- mean(birthwt$ptl)
# The estimand: the probability of a zero count.
pi_true <- exp(-lambda)

n <- 20
n_reps <- 10000

set.seed(1945)
# One row per sample of n counts.
samples <- matrix(rpois(n * n_reps, lambda), nrow = n_reps)
s_sum <- rowSums(samples)

# The crude estimator and its Rao-Blackwellization.
est <- data.frame(
  crude = as.numeric(samples[, 1] == 0),
  rao_blackwell = ((n - 1) / n)^s_sum
)

# Mean with a t interval; variance with a normal-approximation interval,
# using the standard error sqrt((m4 - s^4) / N) of a sample variance.
summarize_estimator <- function(x) {
  tt <- t.test(x)
  v <- var(x)
  m4 <- mean((x - mean(x))^4)
  se_v <- sqrt((m4 - v^2) / length(x))
  data.frame(
    mean = mean(x),
    mean_lower = tt$conf.int[1],
    mean_upper = tt$conf.int[2],
    variance = v,
    var_lower = v - 1.96 * se_v,
    var_upper = v + 1.96 * se_v
  )
}
rb_summary <- do.call(rbind, lapply(est, summarize_estimator))
round(rb_summary, 4)
#>                 mean mean_lower mean_upper variance var_lower var_upper
#> crude         0.8159     0.8083     0.8235   0.1502    0.1454    0.1550
#> rao_blackwell 0.8229     0.8213     0.8245   0.0066    0.0064    0.0068
```

The estimand is $\pi = 0.822$. Both estimators have simulated sample means near it, as Theorem 0.1 says they must. The variance of the Rao–Blackwellized estimator is about 22.7 times smaller than that of the crude estimator, as Theorem 0.2 requires.

```
est_long <- data.frame(
  estimate = c(est$crude, est$rao_blackwell),
  estimator = rep(
    c("crude (first observation is zero)", "Rao-Blackwellized"),
    each = n_reps
  )
)
p <- ggplot(est_long, aes(x = estimate)) +
  geom_histogram(bins = 30) +
  geom_vline(xintercept = pi_true, linetype = "dashed") +
  facet_wrap(~estimator, ncol = 1, scales = "free_y") +
  labs(x = "estimate of the probability of a zero count", y = "count")
if (knitr::is_html_output()) plotly::ggplotly(p) else p
```

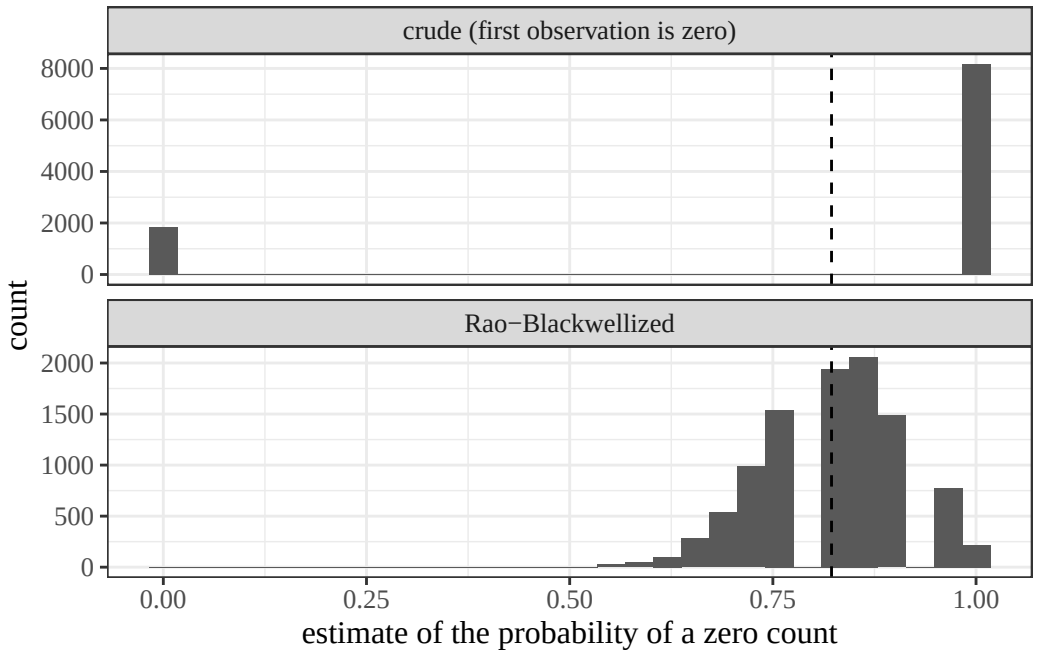


Figure 1: The sampling distribution of the crude and Rao–Blackwellized estimators ($n = 20$, 10,000 samples); the dashed line marks the estimand

> **Remark (Sufficiency is needed)**
>
> *Remark* 0.1 (Sufficiency is needed). If $T(\underset{\sim}{X})$ is not sufficient, then $\mathbb{E}\left[\hat{\theta} \mid T(\underset{\sim}{X})\right]$ may depend on θ, in which case it is not an estimator at all (Casella and Berger [4], Example 7.3.18).

Further reading

- The Fall 2024 course site for UC Berkeley's Stat 210A, *Theoretical Statistics* (Will Fithian) has a course reader on exponential families, sufficiency, completeness and unbiased estimation.

References

[1] Wikipedia contributors, "Statistic — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Statistic

[2] Wikipedia contributors, "Sufficient statistic — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Sufficient_statistic

[3] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A*, vol. 222, no. 594–604, pp. 309–368, 1922, doi: 10.1098/rsta.1922.0009.

[4] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[5] P. R. Halmos and L. J. Savage, "Application of the Radon–Nikodym theorem to the theory of sufficient statistics," *The Annals of Mathematical Statistics*, vol. 20, no. 2, pp. 225–241, 1949, doi: 10.1214/aoms/1177730032.

[6] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–91, 1945.

[7] D. Blackwell, "Conditional expectation and unbiased sequential estimation," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 105–110, 1947, doi: 10.1214/aoms/1177730497.

[8] Wikipedia contributors, "Rao–blackwell theorem — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Rao%E2%80%93Blackwell_theorem

[9] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*. New York: Wiley, 1989.