Probability

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
```

```r
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```r
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

Most of the content in this chapter should be review from UC Davis Epi 202.

1 Core properties of probabilities

1.1 Defining probabilities

Definition 1.1 (Probability measure)**.** A **probability measure**, often denoted Pr() or P(), is a function whose domain is a σ-algebra[a] of possible outcomes, $\mathcal{S}$, and which satisfies the following properties:

1. For any statistical event $A \in \mathcal{S}$, $\Pr(A) \geq 0$.
2. The probability of the union of all outcomes ($\Omega \stackrel{\text{def}}{=} \cup \mathcal{S}$) is 1:

$$\Pr(\Omega) = 1$$

3. The probability of the union of countably many mutually disjoint events $A_1, A_2, ...$ (where $A_i \cap A_j = \emptyset$ for all $i \neq j$) is equal to the sum of their probabilities (*countable additivity* or *sigma-additivity*):

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$$

[a] https://en.wikipedia.org/wiki/%CE%A3-algebra

Property 3 (*countable additivity*) is stronger than *finite additivity*, which only requires

$$\Pr(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \Pr(A_i)$$

for every finite collection of mutually disjoint events. Countable additivity implies finite additivity (set $A_{n+1} = A_{n+2} = \cdots = \emptyset$ in property 3, using $\Pr(\emptyset) = 0$), but not vice versa: there exist set functions that satisfy finite additivity but fail countable additivity (see Wikipedia: Sigma-additive set function — An additive function which is not -additive[1]). Requiring countable additivity enables results such as the continuity of probability (if $A_1 \supseteq A_2 \supseteq \cdots$ with $\bigcap_i A_i = \emptyset$, then $\Pr(A_i) \to 0$) and underpins the Theorem 1.4 for countable partitions.

Theorem 1.1 (Probability of a subset's intersection)**.** *If A and B are statistical events and $A \subseteq B$, then* $\Pr(A \cap B) = \Pr(A)$.

Proof

Proof. Left to the reader for now. □

Theorem 1.2 (An event and its complement sum to 1)**.**

$$\Pr(A) + \Pr(\neg A) = 1$$

Proof

Proof. By properties 2 and 3 of Definition 1.1. □

Corollary 1.1 (Complement rule)**.**

$$\Pr(\neg A) = 1 - \Pr(A)$$

Proof

Proof. By Theorem 1.2 and algebra. □

---

[1] https://en.wikipedia.org/wiki/Sigma-additive_set_function#An_additive_function_which_is_not_%CF%83-additive

Corollary 1.2 (Complement rule in probability (π) notation)**.** *If the probability of an outcome A is $\Pr(A) = \pi$, then the probability that A does not occur is:*

$$\Pr(\neg A) = 1 - \pi$$

Proof

Proof. Using Corollary 1.1:

$$\begin{aligned}\Pr(\neg A) &= 1 - \Pr(A) \\ &= 1 - \pi\end{aligned}$$

□

1.2 Conditional probability

Definition 1.2 (Conditional probability)**.** For two events A and B with $\Pr(B) > 0$, the **conditional probability** of A given B, denoted $\Pr(A \mid B)$, is:

$$\Pr(A \mid B) \stackrel{\text{def}}{=} \frac{\Pr(A \cap B)}{\Pr(B)}$$

Theorem 1.3 (Law of conditional probability)**.** *For any two events A and B with $\Pr(B) > 0$:*

$$\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B)$$

Proof

Proof. Rearranging Definition 1.2:

$$\begin{aligned}\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} && \text{(definition of conditional probability)} \\ \Pr(A \cap B) &= \Pr(A \mid B) \cdot \Pr(B) && \text{(multiply both sides by } \Pr(B))\end{aligned}$$

□

Exm

Example 1.1 (Applying the law of conditional probability)**.** Suppose 30% of adults exercise regularly ($\Pr(E) = 0.30$), and among adults who exercise regularly, 60% have low blood pressure ($\Pr(L \mid E) = 0.60$).
Then the probability that a randomly selected adult both exercises regularly and has low blood pressure is:

$$\begin{aligned}\Pr(L \cap E) &= \Pr(L \mid E) \cdot \Pr(E) \\ &= 0.60 \cdot 0.30 \\ &= 0.18\end{aligned}$$

Theorem 1.4 (Law of total probability)**.** *If $B_1, B_2, \ldots$ is a countable partition of the sample space (i.e., countably many mutually exclusive events whose union is the entire sample space),*

then for any event A:

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \mid B_i) \cdot \Pr(B_i)$$

Proof

Proof. Since $B_1, B_2, \ldots$ partition the sample space, the events $A \cap B_1, A \cap B_2, \ldots$ are mutually exclusive and their union is A. By property 3 of Definition 1.1 (countable additivity), and then by Theorem 1.3:

$$\begin{aligned} \Pr(A) &= \sum_{i=1}^{\infty} \Pr(A \cap B_i) && \text{(countable additivity for partition of } A\text{)} \\ &= \sum_{i=1}^{\infty} \Pr(A \mid B_i) \cdot \Pr(B_i) && \text{(law of conditional probability)} \end{aligned}$$

□

Theorem 1.5 (Bayes' theorem)**.** *For any two events A and B with $\Pr(A) > 0$ and $\Pr(B) > 0$:*

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

Proof

Proof. Apply Definition 1.2 to both $\Pr(A \mid B)$ and $\Pr(B \mid A)$:

$$\begin{aligned} \Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} \end{aligned}$$

The second equality follows from Theorem 1.3 applied to $\Pr(B \cap A) = \Pr(B \mid A) \cdot \Pr(A)$. □

Exm

Example 1.2 (Positive predictive value of a medical test)**.** Suppose a disease test has 99% sensitivity and 99% specificity, and the prevalence of the disease in the population is 7%. Let D be the event "person has the disease" and $+$ be the event "test is positive". Then:

- $\Pr(+ \mid D) = 0.99$ (sensitivity)
- $\Pr(\neg + \mid \neg D) = 0.99$ (specificity), so the false positive rate is $\Pr(+ \mid \neg D) = 1 - 0.99 = 0.01$
- $\Pr(D) = 0.07$ (prevalence)

By Bayes' theorem (Theorem 1.5) and the law of total probability (Theorem 1.4):

$$
\begin{aligned}
\Pr(D \mid +) &= \frac{\Pr(+ \mid D) \cdot \Pr(D)}{\Pr(+)} \\
&= \frac{\Pr(+ \mid D) \cdot \Pr(D)}{\Pr(+ \mid D) \cdot \Pr(D) + \Pr(+ \mid \neg D) \cdot \Pr(\neg D)} \\
&= \frac{0.99 \cdot 0.07}{0.99 \cdot 0.07 + 0.01 \cdot 0.93} \\
&= \frac{0.0693}{0.0693 + 0.0093} \\
&= \frac{0.0693}{0.0786} \\
&\approx 0.88
\end{aligned}
$$

Even with a highly accurate test (99% sensitive and 99% specific), only about 88% of people who test positive actually have the disease, because the disease prevalence is relatively low (7%).

2 Key probability distributions

Some distributions are typically used for outcome models (Table 1); other distributions are typically used for test statistics (Table 2).

Table 1: Distributions typically used for outcome models

Distribution	Uses
Bernoulli	Binary outcomes
Binomial	Sums of Bernoulli outcomes
Poisson	unbounded count outcomes
Geometric	Counts of non-events before an event occurs
Negative binomal	Mixtures of Poisson distributions, counts of non-events until a given number of events occurs
Normal (Gaussian)	Continuous outcomes without a more specific distribution
exponential	Time to event outcomes
Gamma	Time to event outcomes
Weibull	Time to event outcomes
Log-normal	Time to event outcomes

Table 2: Distributions typically used for test statistics

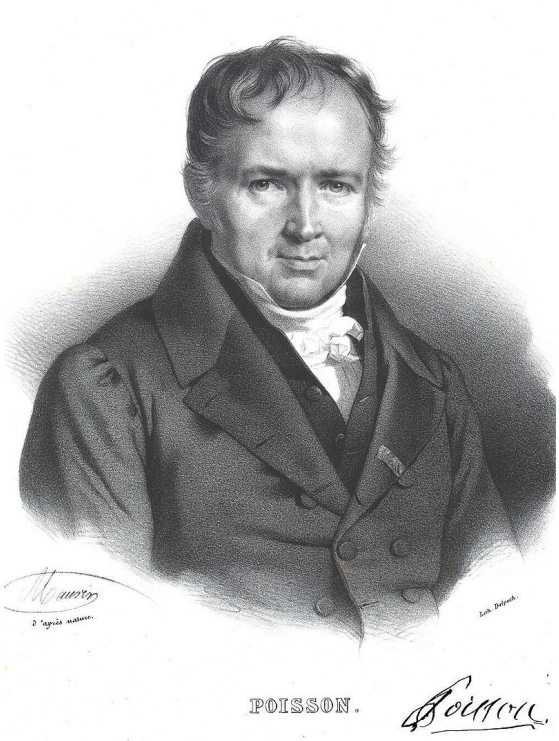
Distribution	Uses
χ^2	Regression comparisons (asymptotic), contingency table independence tests, goodness-of-fit tests
F	Gaussian model comparisons (exact)
Z (standard normal)	Proportions, means, regression coefficients (asymptotic)
T	Means, regression coefficients in Gaussian outcome models (exact)

2.1 The Bernoulli distribution

Definition 2.1 (Bernoulli distribution)**.** The **Bernoulli distribution** family for a random variable X is defined as:

$$\begin{aligned}\Pr(X = x) &\stackrel{\text{def}}{=} 1_{x\in\{0,1\}}\pi^x(1-\pi)^{1-x} \\ &= \begin{cases}\pi, & x = 1 \\ 1-\pi, & x = 0\end{cases}\end{aligned}$$

2.2 The Poisson distribution



(a) Siméon Denis Poisson

(b) Les Poissons[a]

[a] https://youtu.be/UoJxBEQRLd0?t=12

Figure 1: "Les Poissons"

Exercise 2.1. Define the Poisson distribution.

Solution

Solution 2.1.

Def

Definition 2.2 (Poisson distribution).

$$\mathrm{P}(Y = y) \stackrel{\text{def}}{=} \frac{\mu^y e^{-\mu}}{y!}, \quad y \in \mathbb{N} \tag{1}$$

(see Figure 2)

Exercise 2.2. What is the range of possible values for a Poisson distribution?

Solution

Solution 2.2.

$$\mathcal{R}(Y) \stackrel{\text{def}}{=} \{0, 1, 2, \dots\} = \mathbb{N}$$

Theorem 2.1 (CDF of Poisson distribution)**.**

$$\mathrm{P}(Y \leq y) = e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} \tag{2}$$

Proof

Proof. For any $y \geq 0$, the event $\{Y \leq y\}$ is the disjoint union of events $\{Y = j\}$ for all non-negative integers $j \leq \lfloor y \rfloor$. Applying the Poisson PMF (Equation 1) and factoring out the common term $e^{-\mu}$:

$$\begin{aligned} \mathrm{P}(Y \leq y) &= \sum_{j=0}^{\lfloor y \rfloor} \mathrm{P}(Y = j) && \text{(disjoint union of events } Y = j\text{)} \\ &= \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j e^{-\mu}}{j!} && \text{(definition of Poisson PMF)} \\ &= e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} && \text{(factoring out } e^{-\mu} \text{ constant wrt } j\text{)} \end{aligned}$$

□

Exm

Example 2.1 (Example: Computing Poisson cumulative probabilities)**.** For a Poisson random variable $X \sim \mathrm{Pois}(\mu = 2)$, the probability of observing at most 2 events is computed as:

$$\begin{aligned} \mathrm{P}(X \leq 2) &= e^{-2} \sum_{j=0}^{2} \frac{2^j}{j!} && \text{(apply CDF formula with } \mu = 2, y = 2\text{)} \\ &= e^{-2} \left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} \right) && \text{(expand terms for } j = 0, 1, 2\text{)} \\ &= e^{-2}(1 + 2 + 2) && \text{(simplify factorials and powers)} \\ &= 5e^{-2} \approx 0.677 && \text{(evaluate numerical value)} \end{aligned}$$

(see Figure 3)

```
library(dplyr)
pois_dists <- tibble(
  mu = c(0.5, 1, 2, 5, 10, 20)
) |>
  reframe(
    .by = mu,
    x = 0:30
  ) |>
  mutate(
    `P(X = x)` = dpois(x, lambda = mu),
    `P(X <= x)` = ppois(x, lambda = mu),
    mu = factor(mu)
  )

library(ggplot2)
library(latex2exp)

plot0 <- pois_dists |>
  ggplot(
    aes(
      x = x,
      y = `P(X = x)`,
      fill = mu,
      col = mu
    )
  ) +
```

```
  theme(legend.position = "bottom") +
  labs(
    fill = latex2exp::TeX("$\\mu$"),
    col = latex2exp::TeX("$\\mu$"),
    y = latex2exp::TeX("$\\Pr_{\\mu}(X = x)$")
  )

plot1 <- plot0 +
  geom_segment(yend = 0) +
  facet_wrap(~mu)

print(plot1)
```

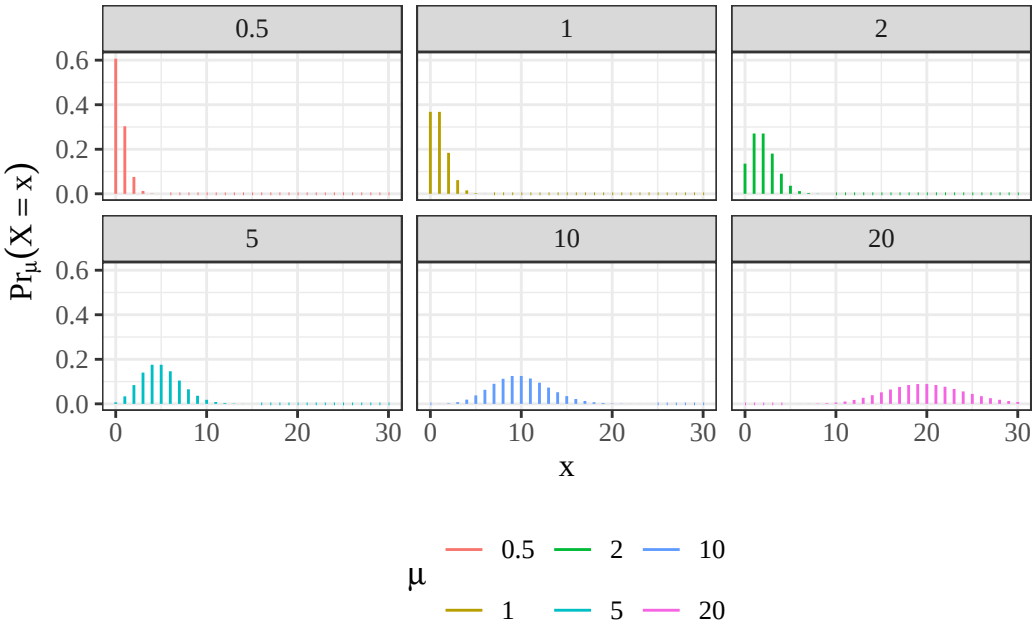


Figure 2: Poisson PMFs, by mean parameter μ

```
library(ggplot2)

plot2 <-
  plot0 +
  geom_step(alpha = 0.75) +
  aes(y = `P(X <= x)`) +
  labs(y = latex2exp::TeX("$\\Pr_{\\mu}(X \\leq x)$"))

print(plot2)
```

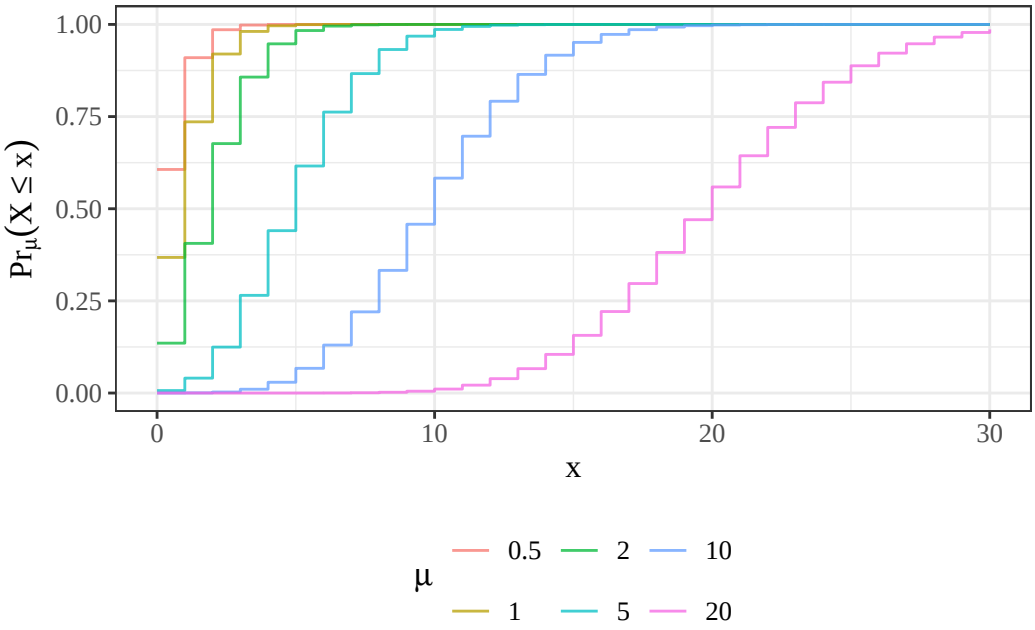


Figure 3: Poisson CDFs

Exercise 2.3 (Poisson distribution functions)**.** Let $X \sim \text{Pois}(\mu = 3.75)$.
Compute:

- $\text{P}(X = 4|\mu = 3.75)$
- $\text{P}(X \leq 7|\mu = 3.75)$
- $\text{P}(X > 5|\mu = 3.75)$

Solution

Solution.

- $\text{P}(X = 4) = 0.19378$
- $\text{P}(X \leq 7) = 0.962379$
- $\text{P}(X > 5) = 0.177117$

Theorem 2.2 (Properties of the Poisson distribution)**.** *If* $X \sim \text{Pois}(\mu)$*, then:*

- $\text{E}[X] = \mu$
- $\text{Var}(X) = \mu$
- $\text{P}(X = x) = \frac{\mu}{x}\,\text{P}(X = x - 1)$
- *For* $x < \mu$, $\text{P}(X = x) > \text{P}(X = x - 1)$
- *For* $x = \mu$, $\text{P}(X = x) = \text{P}(X = x - 1)$
- *For* $x > \mu$, $\text{P}(X = x) < \text{P}(X = x - 1)$
- $\arg\max_x \text{P}(X = x) = \lfloor \mu \rfloor$

Exercise 2.4. Prove Theorem 2.2.

Solution

Solution.

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{x=0}^{\infty} x \cdot \mathrm{P}(X = x) && \text{(definition of expected value)} \\
&= 0 \cdot \mathrm{P}(X = 0) + \sum_{x=1}^{\infty} x \cdot \mathrm{P}(X = x) && \text{(separate } x = 0 \text{ term)} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\mu^x e^{-\mu}}{x!} && \text{(substitute Poisson PMF)} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\mu^x e^{-\mu}}{x \cdot (x-1)!} && \text{(definition of factorial } x!) \\
&= \sum_{x=1}^{\infty} \frac{\mu^x e^{-\mu}}{(x-1)!} && \text{(cancel factor of } x) \\
&= \mu \cdot \sum_{x=1}^{\infty} \frac{\mu^{x-1} e^{-\mu}}{(x-1)!} && \text{(factor out one power of } \mu) \\
&= \mu \cdot \sum_{y=0}^{\infty} \frac{\mu^y e^{-\mu}}{y!} && \text{(change index variable } y \stackrel{\text{def}}{=} x - 1) \\
&= \mu \cdot 1 && \text{(PMF sums to 1 over state space)} \\
&= \mu && \text{(simplify)}
\end{aligned}
$$

See also https://statproofbook.github.io/P/poiss-mean.
For the variance, see https://statproofbook.github.io/P/poiss-var.

Accounting for exposure

Definition 2.3 (Exposure magnitude)**.** For many count outcomes, there is some sense of an **exposure magnitude**, such as **population size**, or **duration of observation**, which multiplicatively rescales the expected (mean) count.

Exercise 2.5. What are some examples of exposure magnitudes?

Solution

Solution.

Table 3: Examples of exposure units

outcome	exposure units
disease incidence	number of individuals exposed; time at risk
car accidents	miles driven
worksite accidents	person-hours worked
population size	size of habitat

Exposure units are similar to the number of trials in a binomial distribution, but **in non-binomial count outcomes, there can be more than one event per unit of exposure**.

We can use t to represent continuous-valued exposures/observation durations, and n to represent discrete-valued exposures.

Definition 2.4 (Event rate)**.**
For a count outcome Y with exposure magnitude t, the **event rate** (denoted λ) is defined as the mean of Y divided by the exposure magnitude. This mathematical relationship between mean and exposure magnitude is:

$$\mu \stackrel{\text{def}}{=} \mathrm{E}[Y|T=t]$$

$$\lambda \stackrel{\text{def}}{=} \frac{\mu}{t} \tag{3}$$

Event rate is somewhat analogous to odds in binary outcome models; it typically serves as an intermediate transformation between the mean of the outcome and the linear component of the model. However, in contrast with the odds function, the transformation $\lambda = \mu/t$ is *not* considered part of the Poisson model's link function, and it treats the exposure magnitude covariate differently from the other covariates.

Theorem 2.3 (Transformation function from event rate to mean)**.** *For a count variable with mean μ, event rate λ, and exposure magnitude t:*

$$\mu = \lambda \cdot t \tag{4}$$

Proof

Proof.

$$\begin{aligned} \lambda &\stackrel{\text{def}}{=} \frac{\mu}{t} && \text{(definition of event rate } @eq-def-event-rate) \\ \mu &= \lambda \cdot t && \text{(multiply both sides by } t > 0) \end{aligned}$$

□

Exm

Example 2.2 (Example: Calculating expected counts from event rates)**.** Suppose a city records a disease event rate of $\lambda = 0.05$ cases per person-year. For a subpopulation with an exposure magnitude of $t = 100$ person-years, the expected count of cases is:

$$\begin{aligned} \mu &= \lambda \cdot t && \text{(apply transformation formula } @eq-lambda-to-mu) \\ &= 0.05 \times 100 && \text{(substitute } \lambda = 0.05 \text{ and } t = 100) \\ &= 5 \text{ cases} && \text{(evaluate expected count)} \end{aligned}$$

Solution

Solution. Start from definition of event rate and use algebra to solve for μ.

Equation 4 is analogous to the inverse-odds function for binary variables.

Theorem 2.4 (No exposure means no expected events)**.** *When the exposure magnitude is 0, there is no opportunity for events to occur:*

$$\mathrm{E}[Y|T=0] = 0$$

Proof

Proof.

$$\begin{aligned} \mathrm{E}[Y \mid T = 0] &= \lambda \cdot 0 && \text{(apply transformation } @eq - lambda - to - mu \text{ at } t = 0) \\ &= 0 && \text{(multiplication by zero)} \end{aligned}$$

□

Exm

Example 2.3 (Example: Zero exposure time)**.** If a subject is observed for $t = 0$ person-years, no follow-up time has elapsed, so the expected number of incident events is $\mathrm{E}[Y \mid T = 0] = 0$.

Important

The exposure magnitude, T, is *similar* to a covariate in linear or logistic regression. However, there is an important difference: in count regression, **there is no intercept corresponding to** $\mathrm{E}[Y|T = 0]$. In other words, this model assumes that if there is no exposure, there can't be any events.

Theorem 2.5 (Exposure is additive on the log scale)**.** *If $\mu = \lambda \cdot t$, then:*

$$\log \mu = \log \lambda + \log t$$

Proof

Proof.

$$\begin{aligned} \log \mu &= \log(\lambda \cdot t) && \text{(substitute } \mu = \lambda \cdot t \text{ from } @eq - lambda - to - mu) \\ &= \log \lambda + \log t && \text{(logarithm product rule)} \end{aligned}$$

□

Exm

Example 2.4 (Example: Log-linear representation of expected counts)**.** If a clinic sees an event rate of $\lambda = 0.02$ events/day and $t = 30$ days of observation:

$$\begin{aligned} \log \mu &= \log(0.02) + \log(30) && \text{(apply log-scale formula)} \\ &= -3.912 + 3.401 && \text{(evaluate natural logarithms)} \\ &= -0.511 && \text{(sum terms)} \end{aligned}$$

Exponentiating yields $\mu = \exp\{-0.511\} \approx 0.60$ expected events.

Definition 2.5 (Offset)**.** When the linear component of a model involves a term without an unknown coefficient, that term is called an **offset**.

Theorem 2.6 (Sum of independent Poisson random variables)**.** *If X and Y are independent Poisson random variables with means μ_X and μ_Y, their sum, $Z = X + Y$, is also a Poisson random variable, with mean $\mu_Z = \mu_X + \mu_Y$.*

Proof

Proof. Using the probability-generating function or PMF convolution for independent non-negative integer random variables:

$$\begin{aligned}
\mathrm{P}(Z=z) &= \sum_{k=0}^{z} \mathrm{P}(X=k)\,\mathrm{P}(Y=z-k) && \text{(independence and convolution formula)}\\
&= \sum_{k=0}^{z} \frac{\mu_X^k e^{-\mu_X}}{k!}\,\frac{\mu_Y^{z-k} e^{-\mu_Y}}{(z-k)!} && \text{(substitute Poisson PMFs)}\\
&= \frac{e^{-(\mu_X+\mu_Y)}}{z!} \sum_{k=0}^{z} \frac{z!}{k!(z-k)!}\mu_X^k \mu_Y^{z-k} && \text{(factor out } e^{-(\mu_X+\mu_Y)}/z! \text{ and multiply by } z!/z!)\\
&= \frac{e^{-(\mu_X+\mu_Y)}}{z!}(\mu_X+\mu_Y)^z && \text{(binomial theorem)}
\end{aligned}$$

This probability expression matches the PMF of a $\mathrm{Pois}(\mu_X + \mu_Y)$ random variable (see also https://web.stanford.edu/class/archive/cs/cs109/cs109.1206/lectureNotes/LN12_independent_rvs.pdf, Example 3). □

Exm

Example 2.5 (Example: Aggregating independent region counts)**.** Suppose Region A records $X \sim \mathrm{Pois}(\mu_X = 12)$ cases and Region B records $Y \sim \mathrm{Pois}(\mu_Y = 18)$ cases independently. The combined total count $Z = X + Y$ follows a Poisson distribution:

$$\begin{aligned}
Z &\sim \mathrm{Pois}(\mu_X + \mu_Y) && \text{(apply sum theorem } @thm - sum - pois)\\
&= \mathrm{Pois}(12+18) && \text{(substitute region means)}\\
&= \mathrm{Pois}(30) && \text{(evaluate sum)}
\end{aligned}$$

2.3 The Negative-Binomial distribution

Definition 2.6 (Negative binomial distribution)**.**

$$\mathrm{P}(Y=y) \stackrel{\text{def}}{=} \frac{\mu^y}{y!} \cdot \frac{\Gamma(\rho+y)}{\Gamma(\rho)\cdot(\rho+\mu)^y} \cdot \left(1+\frac{\mu}{\rho}\right)^{-\rho}$$

where ρ is an overdispersion parameter and $\Gamma(x) = (x-1)!$ for integers x.

You don't need to memorize or understand this expression.

As $\rho \to \infty$, the second term converges to 1 and the third term converges to $\exp\{-\mu\}$, which brings us back to the Poisson distribution.

Theorem 2.7 (Mean and variance of the negative binomial distribution)**.** *If* $Y \sim \mathrm{NegBin}(\mu, \rho)$*, then:*

- $\mathrm{E}[Y] = \mu$
- $\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\rho} > \mu$

2.4 Weibull Distribution

$$
\begin{aligned}
p(t) &= \alpha\lambda x^{\alpha-1}\mathrm{e}^{-\lambda x^{\alpha}} \\
\lambda(t) &= \alpha\lambda x^{\alpha-1} \\
\mathrm{S}(t) &= \mathrm{e}^{-\lambda x^{\alpha}} \\
E(T) &= \Gamma(1+1/\alpha)\cdot\lambda^{-1/\alpha}
\end{aligned}
$$

When $\alpha = 1$, the Weibull distribution reduces to the exponential distribution. When $\alpha > 1$ the hazard is increasing and when $\alpha < 1$ the hazard is decreasing. This property provides more flexibility than the exponential.

We will see more of this distribution later.

3 Characteristics of probability distributions

3.1 Probability density function

Definition 3.1 (probability density)**.** If X is a continuous random variable, then the **probability density** of X at value x, denoted $f(x)$, $f_X(x)$, $\mathrm{p}(x)$, $\mathrm{p}_X(x)$, or $\mathrm{p}(X = x)$, is defined as the limit of the probability (mass) that X is in an interval around x, divided by the width of that interval, as that width reduces to 0.

$$f(x) \stackrel{\text{def}}{=} \lim_{\Delta\to 0} \frac{\mathrm{P}(X \in [x, x+\Delta])}{\Delta}$$

See also Rothman et al. (2021) (Chapter 22, p. 535) and https://en.wikipedia.org/wiki/Probability_density_function#Formal_definition

---

Definition 3.2 (Cumulative distribution function (CDF))**.** For a random variable X, its population CDF is

$$F(t) \stackrel{\text{def}}{=} \Pr(X \le t), \quad t \in \mathbb{R}.$$

Definition 3.3 (Quantile function (population inverse CDF))**.** For a random variable X with cumulative distribution function (CDF) F, its population quantile function (generalized inverse of F) is

$$Q(p) \stackrel{\text{def}}{=} \inf\{t : F(t) \ge p\}, \quad 0 < p \le 1.$$

Theorem 3.1 (Density function is derivative of CDF)**.** *The density function $f(t)$ or $\mathrm{p}(T = t)$ for a random variable T at value t is equal to the derivative of the cumulative probability function $F(t) \stackrel{def}{=} P(T \le t)$; that is:*

$$f(t) \stackrel{def}{=} \frac{\partial}{\partial t} F(t)$$

Theorem 3.2 (Density functions integrate to 1)**.** *For any density function $f(x)$,*

$$\int_{x\in\mathcal{R}(X)} f(x)dx = 1$$

3.2 Hazard function

Definition 3.4 (Hazard function, hazard rate, hazard rate function)**.**
The **hazard function**, **hazard rate**, **hazard rate function**, for a random variable T at value t, typically denoted as $\mathrm{h}(t)$ [2] or $\lambda(t)$, [3] is the conditional density[a] of T at t, given $T \geq t$. Specifically, the hazard function is defined as:

$$\lambda(t) \stackrel{\text{def}}{=} \mathrm{p}(T = t | T \geq t)$$

If T represents the time at which an event occurs, then $\lambda(t)$ is the probability that the event occurs at time t, given that it has not occurred prior to time t.
The name "hazard" carries a connotation that the event is undesirable — death, relapse, equipment failure, etc. When the event in question is neutral or desirable (recovery, conception, graduation, response to treatment), the same quantity $\lambda(t)$ is often called the **event incidence rate** instead. This terminology is parallel to the convention that conditional probabilities of undesirable events are called **risks**, while the same conditional probabilities for neutral/desirable events are simply called **probabilities**. The math is identical; only the name changes with the valence of the event.

[a] probability.qmd#def-pdf

Table 4: Probability distribution functions

Name	Symbols	Definition
Probability density function (PDF)	$\mathrm{f}(t), \mathrm{p}(t)$	$\mathrm{p}(T = t)$
Cumulative distribution function (CDF)	$\mathrm{F}(t), \mathrm{P}(t)$	$\mathrm{P}(T \leq t)$
Survival function	$\mathrm{S}(t), \bar{\mathrm{F}}(t)$	$\mathrm{P}(T > t)$
Hazard function	$\lambda(t), \mathrm{h}(t)$	$\mathrm{p}(T = t \mid T \geq t)$
Cumulative hazard function	$\Lambda(t), \mathrm{H}(t)$	$\int_{u=-\infty}^{t} \lambda(u)du$
Log-hazard function	$\eta(t)$	$\log\{\lambda(t)\}$

$$\mathrm{f}(t) \underset{\mathrm{S}(t)\lambda(t)}{\overset{-S'(t)}{\longleftarrow}} \mathrm{S}(t) \xleftarrow{\exp\{-\Lambda(t)\}} \Lambda(t) \xleftarrow{\int_{u=0}^{t} \lambda(u)du} \lambda(t) \xleftarrow{\exp\{\eta(t)\}} \eta(t)$$

$$\mathrm{f}(t) \xrightarrow[\int_{u=t}^{\infty} \mathrm{f}(u)du]{\mathrm{f}(t)/\lambda(t)} \mathrm{S}(t) \xrightarrow[-\log \mathrm{S}(t)]{} \Lambda(t) \xrightarrow[\Lambda'(t)]{} \lambda(t) \xrightarrow[\log\{\lambda(t)\}]{} \eta(t)$$

3.3 Expectation

Definition 3.5 (Expectation, expected value, population mean)**.** The **expectation**, **expected value**, or **population mean** of a *continuous* random variable X, denoted $\mathrm{E}[X]$, $\mu(X)$, or μ_X, is the weighted mean of X's possible values, weighted by the probability density function of those values:

$$\mathrm{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{R}(X)} x \cdot \mathrm{p}(X = x)dx$$

The **expectation**, **expected value**, or **population mean** of a *discrete* random variable X, denoted $\mathrm{E}[X]$, $\mu(X)$, or μ_X, is the mean of X's possible values, weighted by the probability mass function of those values:

[3] for example in Dobson and Barnett (2018), Vittinghoff et al. (2012), Klein and Moeschberger (2003), and Kleinbaum and Klein (2012)
[3] for example, in Rothman et al. (2021) and Kalbfleisch and Prentice (2011)

$$\mathrm{E}[X] \stackrel{\text{def}}{=} \sum_{x \in \mathcal{R}(X)} x \cdot \mathrm{P}(X = x)$$

(c.f. https://en.wikipedia.org/wiki/Expected_value)

Theorem 3.3 (Expectation of the Bernoulli distribution)**.** *The expectation of a Bernoulli random variable with parameter π is:*

$$\mathrm{E}[X] = \pi$$

Proof

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x \in \mathcal{R}(X)} x \cdot \mathrm{P}(X = x) && \text{(definition of expectation for discrete r.v.)} \\
&= \sum_{x \in \{0,1\}} x \cdot \mathrm{P}(X = x) && \text{(range of Bernoulli r.v. is } \{0,1\}) \\
&= (0 \cdot \mathrm{P}(X = 0)) + (1 \cdot \mathrm{P}(X = 1)) && \text{(expand sum over } x = 0 \text{ and } x = 1) \\
&= (0 \cdot (1 - \pi)) + (1 \cdot \pi) && \text{(substitute Bernoulli PMF values)} \\
&= 0 + \pi && \text{(multiplication by 0 and 1)} \\
&= \pi && \text{(addition of 0)}
\end{aligned}$$

□

Theorem 3.4 (Expectation of time-to-event variables)**.** *If T is a non-negative random variable, then:*

$$\mathrm{E}[T] = \int_{t=0}^{\infty} \mathrm{S}(t)\, dt$$

Proof

Proof. We prove the continuous case, in which T has a density f. The result follows from applying Tonelli's theorem (condition (a) of Fubini–Tonelli[a]) to the function $g(t, u) = \mathrm{f}(u) \cdot \mathbb{1}(0 \le t \le u)$ on the product space $[0, \infty) \times [0, \infty)$: g is nonnegative everywhere and vanishes outside the (unbounded) triangular region $D = \{(t, u) : 0 \le t \le u < \infty\}$, so the iterated integrals over D are exchangeable.
Since $\mathrm{f}(u) \ge 0$, condition (a) of Fubini–Tonelli[b] (the nonnegative case, **Tonelli's theorem**) applies, and we may exchange the order of integration over D:

$$
\begin{aligned}
\mathrm{E}[T] &= \int_{u=0}^{\infty} u \, \mathrm{f}(u) \, du \\
&= \int_{u=0}^{\infty} \left(\int_{t=0}^{u} 1 \, dt \right) \mathrm{f}(u) \, du \\
&= \int_{u=0}^{\infty} \int_{t=0}^{u} \mathrm{f}(u) \, dt \, du \\
&= \int_{t=0}^{\infty} \int_{u=t}^{\infty} \mathrm{f}(u) \, du \, dt \\
&= \int_{t=0}^{\infty} \mathrm{P}(T > t) \, dt \\
&= \int_{t=0}^{\infty} \mathrm{S}(t) \, dt.
\end{aligned}
$$

See (Soch 2020) for an alternative presentation of this proof. □

[a] math-prereqs.qmd#thm-fubini-tonelli
[b] math-prereqs.qmd#thm-fubini-tonelli

Exm

Example 3.1 (Mean of an exponential random variable via survival function)**.** Let $T \sim \text{Exponential}(\lambda)$, so $\mathrm{S}(t) = \mathrm{e}^{-\lambda t}$ for $t \geq 0$. By Theorem 3.4:

$$
\begin{aligned}
\mathrm{E}[T] &= \int_{0}^{\infty} \mathrm{S}(t) \, dt \\
&= \int_{0}^{\infty} \mathrm{e}^{-\lambda t} \, dt \\
&= \left[-\frac{1}{\lambda} \mathrm{e}^{-\lambda t} \right]_{0}^{\infty} \\
&= \frac{1}{\lambda},
\end{aligned}
$$

confirming the standard result $\mathrm{E}[T] = 1/\lambda$.

Theorem 3.5 (Law of the Unconscious Statistician (LOTUS))**.** ***Discrete case.*** *For any function g of a* discrete *random variable X:*

$$
\mathrm{E}[g(X)] = \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathrm{P}(X = x)
$$

Continuous case. *For any function g of a* continuous *random variable X with density* $\mathrm{p}(X = x)$*:*

$$
\mathrm{E}[g(X)] = \int_{x \in \mathcal{R}(X)} g(x) \cdot \mathrm{p}(X = x) \, dx
$$

Proof

Proof. We prove the discrete case.
Let $Y = g(X)$. By Definition 3.5 applied to Y:

$$
\begin{aligned}
\mathrm{E}[g(X)] &= \mathrm{E}[Y] && \text{(substitution } Y = g(X)\text{)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y) && \text{(definition of expectation for discrete r.v.)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(g(X) = y) && \text{(substitute } Y = g(X) \text{ in probability expression)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{\substack{x \in \mathcal{R}(X) \\ g(x) = y}} \mathrm{P}(X = x) && \text{(law of total probability over } \{x : g(x) = y\}\text{)} \\
&= \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathrm{P}(X = x) && \text{(rearrange double sum grouping by } g(x)\text{)}
\end{aligned}
$$

where the last equality follows by rearranging the double sum, grouping each term x by its image $y = g(x)$.

The continuous case is the density-weighted analogue of this argument, but a fully rigorous proof needs the general change-of-variables theorem for integrals against a pushforward measure — approximating g by simple functions and passing to the limit — which is beyond Epi 204's scope (Billingsley 1995; Gut 2013; Casella and Berger 2002; Wikipedia contributors 2026). □

LOTUS says that to compute $\mathrm{E}[g(X)]$, we do not need to first find the distribution of $g(X)$; we can compute the expectation directly using the distribution of X.

Exm

Example 3.2 (Expected value of X^2 for a Bernoulli variable)**.** Let $X \sim \mathrm{Ber}(\pi)$. By LOTUS (Theorem 3.5, discrete case):

$$
\begin{aligned}
\mathrm{E}[X^2] &= \sum_{x \in \{0,1\}} x^2 \cdot \mathrm{P}(X = x) \\
&= 0^2 \cdot \mathrm{P}(X = 0) + 1^2 \cdot \mathrm{P}(X = 1) \\
&= 0^2 \cdot (1 - \pi) + 1^2 \cdot \pi \\
&= 0 + \pi \\
&= \pi
\end{aligned}
$$

Exm

Example 3.3 (Expected value of X^2 for a Uniform(0,1) variable)**.** Let $X \sim \mathrm{Uniform}(0, 1)$, so $\mathrm{p}(X = x) = 1$ for $x \in [0, 1]$. By LOTUS (Theorem 3.5, continuous case):

$$
\begin{aligned}
\mathrm{E}[X^2] &= \int_0^1 x^2 \cdot \mathrm{p}(X = x)\, dx && \text{(LOTUS, continuous case)} \\
&= \int_0^1 x^2 \cdot 1\, dx && (\,\mathrm{p}(X = x) = 1 \text{ on } [0, 1]) \\
&= \left[\frac{x^3}{3}\right]_0^1 && \text{(antiderivative of } x^2\text{)} \\
&= \frac{1}{3}. && \text{(evaluate at the bounds)}
\end{aligned}
$$

3.3.1 Conditional distributions and expectations

Definition 3.6 (Conditional probability mass function)**.** Let X and Y be jointly distributed discrete random variables. The **conditional probability mass function** of Y given $X = x$ (for values of x with $\mathrm{P}(X = x) > 0$) is:

$$\mathrm{P}(Y = y \mid X = x) \stackrel{\text{def}}{=} \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(X = x)}$$

Exm

Example 3.4 (Conditional PMF from a vaccine trial)**.** The `vaccine` dataset in the `dobson` package (Dobson and Barnett (2018), Table 9.6) records the response to treatment in a vaccine trial: X is treatment group (placebo or vaccine) and Y is response severity (small, moderate, or large).

```
library(dobson)
library(dplyr)
data(vaccine)
response_levels <- c("small", "moderate", "large")
vaccine_tab <- vaccine |>
  mutate(response = factor(response, levels = response_levels)) |>
  xtabs(frequency ~ treatment + response, data = _)
pander::pander(as.data.frame.matrix(vaccine_tab))
```

	small	moderate	large
placebo	25	8	5
vaccine	6	18	11

The joint PMF of (X, Y) is the table of frequencies divided by the total $n = 73$ trial participants. The marginal probability $\mathrm{P}(X = \text{placebo})$ is:

$$\begin{aligned}\mathrm{P}(X = \text{placebo}) &= \mathrm{P}(X = \text{placebo}, Y = \text{small}) + \mathrm{P}(X = \text{placebo}, Y = \text{moderate}) + \mathrm{P}(X = \text{placebo}, Y = \text{large}) \\ &= \tfrac{25}{73} + \tfrac{8}{73} + \tfrac{5}{73} \\ &= \tfrac{38}{73}\end{aligned}$$

By Definition 3.6, the conditional PMF of Y given $X = \text{placebo}$ is:

$$\begin{aligned}\mathrm{P}(Y = \text{small} \mid X = \text{placebo}) &= \frac{\mathrm{P}(X = \text{placebo}, Y = \text{small})}{\mathrm{P}(X = \text{placebo})} \\ &= \frac{25/73}{38/73} \\ &= \frac{25}{38}\end{aligned}$$

Definition 3.7 (Conditional probability density function)**.** Let X and Y be jointly distributed continuous random variables with joint density $\mathrm{p}(X = x, Y = y)$ and marginal density $\mathrm{p}(X = x)$. The **conditional probability density function** of Y given $X = x$ (for values of x with $\mathrm{p}(X = x) > 0$) is:

$$\mathrm{p}(Y = y \mid X = x) \stackrel{\text{def}}{=} \frac{\mathrm{p}(X = x, Y = y)}{\mathrm{p}(X = x)}$$

Exm

Example 3.5 (Conditional PDF from a bivariate normal model of birthweight data)**.** The `birthweight` dataset in the `dobson` package (Dobson and Barnett (2018), Table 2.3) records gestational age (weeks) and birthweight (grams) for 12 boys and 12 girls. Let X be gestational age and Y be birthweight, pooling both sexes into $n = 24$ observations.

```
library(dobson)
data(birthweight)
ga <- c(birthweight[["boys gestational age"]],
        birthweight[["girls gestational age"]])
wt <- c(birthweight[["boys weight"]], birthweight[["girls weight"]])
mu_x <- mean(ga)
sigma_x <- sd(ga)
mu_y <- mean(wt)
sigma_y <- sd(wt)
rho <- cor(ga, wt)
beta <- cov(ga, wt) / var(ga)
intercept <- mu_y - beta * mu_x
x0 <- 40
marg_dens_x0 <- dnorm(x0, mean = mu_x, sd = sigma_x)
cond_mean <- intercept + beta * x0
cond_sd <- sigma_y * sqrt(1 - rho^2)
pander::pander(data.frame(
  parameter = c("mu_x", "sigma_x", "mu_y", "sigma_y", "rho"),
  estimate = round(c(mu_x, sigma_x, mu_y, sigma_y, rho), 4)
))
```

parameter	estimate
mu_x	38.54
sigma_x	1.817
mu_y	2,968
sigma_y	282.1
rho	0.7443

Modeling (X, Y) as bivariate normal with parameters set equal to these sample moments, the joint density is (a standard result; e.g. Casella and Berger (2002)):

$$\mathrm{p}(X = x,\, Y = y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]}$$

A further standard fact about the bivariate normal (Casella and Berger (2002)) is that the marginal distribution of X is $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. At $x = 40$ weeks (a full-term pregnancy), $\mu_X = 38.5417$ and $\sigma_X = 1.8173$, so:

$$\begin{aligned}
\mathrm{p}(X = 40) &= \frac{1}{\sigma_X\sqrt{2\pi}}\mathrm{e}^{-\frac{(40-\mu_X)^2}{2\sigma_X^2}} \\
&= \frac{1}{1.8173\sqrt{2\pi}}\mathrm{e}^{-\frac{(40-38.5417)^2}{2(1.8173)^2}} \\
&\approx 0.1591
\end{aligned}$$

By Definition 3.7, dividing the joint density by this marginal density and simplifying the exponent (completing the square in y; Casella and Berger (2002)) gives the conditional PDF of Y given $X = 40$, which is itself normal with mean shifted along the regression line and variance reduced by a factor of $1 - \rho^2$:

$$\begin{aligned}
\mathrm{p}(Y = y \mid X = 40) &= \frac{\mathrm{p}(X = 40,\, Y = y)}{\mathrm{p}(X = 40)} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(40-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(40-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right] + \frac{(40-\mu_X)^2}{2\sigma_X^2}} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}}\mathrm{e}^{-\frac{\rho^2(40-\mu_X)^2}{2\sigma_X^2(1-\rho^2)} + \frac{\rho(40-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y(1-\rho^2)} - \frac{(y-\mu_Y)^2}{2\sigma_Y^2(1-\rho^2)}} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}}\mathrm{e}^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[\rho^2\frac{\sigma_Y^2}{\sigma_X^2}(40-\mu_X)^2 - 2\rho\frac{\sigma_Y}{\sigma_X}(40-\mu_X)(y-\mu_Y) + (y-\mu_Y)^2\right]} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}}\mathrm{e}^{-\frac{\left(y - \left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(40-\mu_X)\right]\right)^2}{2\sigma_Y^2(1-\rho^2)}}
\end{aligned}$$

The first equality substitutes the joint and marginal densities and combines their prefactors and exponents; the second equality combines the two $(40-\mu_X)^2$ terms using $1 - \frac{1}{1-\rho^2} = \frac{-\rho^2}{1-\rho^2}$; the third equality factors $-\frac{1}{2\sigma_Y^2(1-\rho^2)}$ out of the bracket (multiplying through confirms it reproduces the previous line); and the fourth equality completes the square in y, since expanding $\left(y - \left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(40-\mu_X)\right]\right)^2$ reproduces exactly the bracketed quadratic in the third equality.
So $Y \mid X = 40 \sim \mathrm{N}(3136.15,\, 188.37^2)$: the conditional mean, 3136.15 g, is exactly the fitted regression line's prediction at $x = 40$ ($-1484.9846 + 115.5283 \times 40 = 3136.15$), matching R's `lm(wt ~ ga)` fit directly:

```
pander::pander(coef(lm(wt ~ ga)))
```

(Intercept)	ga
-1,485	115.5

Definition 3.8 (Conditional expectation)**. Discrete case.** Let X and Y be jointly distributed discrete random variables. The **conditional expectation** of Y given $X = x$, using Definition 3.6, is:

$$\mathrm{E}[Y \mid X = x] \overset{\text{def}}{=} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y \mid X = x)$$

Continuous case. Let X and Y be jointly distributed continuous random variables. The **conditional expectation** of Y given $X = x$, using Definition 3.7, is:

$$\mathrm{E}[Y \mid X = x] \overset{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} y \cdot \mathrm{p}(Y = y \mid X = x)\, dy$$

Exm

Example 3.6 (Conditional expectation from real trial and birthweight data)**. Discrete case.** Continuing Example 3.4, score the vaccine trial's response levels small = 1, moderate = 2, large = 3. The conditional PMF of Y given $X =$ placebo (from Example 3.4) is:

$$\begin{aligned}
\mathrm{P}(Y = \text{small} \mid X = \text{placebo}) &= \tfrac{25}{38} \\
\mathrm{P}(Y = \text{moderate} \mid X = \text{placebo}) &= \tfrac{8}{38} \\
\mathrm{P}(Y = \text{large} \mid X = \text{placebo}) &= \tfrac{5}{38}
\end{aligned}$$

so:

$$\begin{aligned}
\mathrm{E}[Y \mid X = \text{placebo}] &= 1 \cdot \tfrac{25}{38} + 2 \cdot \tfrac{8}{38} + 3 \cdot \tfrac{5}{38} \\
&= \frac{25 + 16 + 15}{38} \\
&= \frac{56}{38} \\
&= \frac{28}{19} \\
&\approx 1.47
\end{aligned}$$

Continuous case. Continuing Example 3.5, $Y \mid X = 40 \sim \mathrm{N}(3136.15,\, 188.37^2)$. Since the mean of a normal random variable is its own location parameter, integrating y against this conditional density directly gives:

$$\begin{aligned}
\mathrm{E}[Y \mid X = 40] &= \int_{-\infty}^{\infty} y \cdot \mathrm{p}(Y = y \mid X = 40)\, dy \\
&= 3136.15 \text{ g}
\end{aligned}$$

matching the fitted regression line's prediction at $x = 40$ weeks exactly, as expected since the conditional mean of a bivariate normal *is* the linear regression of Y on X.

Definition 3.9 (Conditional expectation: mixed case)**.** Suppose exactly one of X, Y is discrete and the other is continuous. Write $\mathrm{p}(X = x,\, Y = y)$ for their **joint density-mass function**: a probability density in the continuous variable and a probability mass in the discrete variable. X **discrete,** Y **continuous.** Here $\mathrm{p}(X = x,\, Y = y)$ is, for each fixed x, a probability density in y, scaled so that $\int_y \mathrm{p}(X = x,\, Y = y)\, dy = \mathrm{P}(X = x)$. The conditional PDF of Y given $X = x$ (for values of x with $\mathrm{P}(X = x) > 0$) is:

$$\mathrm{p}(Y = y \mid X = x) \overset{\text{def}}{=} \frac{\mathrm{p}(X = x,\, Y = y)}{\mathrm{P}(X = x)}$$

and the conditional expectation of Y given $X = x$ is:

$$\mathrm{E}[Y \mid X = x] \overset{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} y \cdot \mathrm{p}(Y = y \mid X = x)\, dy$$

X continuous, Y discrete. Here $\mathrm{p}(X = x, Y = y)$ is, for each fixed y, a probability density in x, scaled so that $\sum_y \mathrm{p}(X = x, Y = y) = \mathrm{p}(X = x)$. The conditional PMF of Y given $X = x$ (for values of x with $\mathrm{p}(X = x) > 0$) is:

$$\mathrm{P}(Y = y \mid X = x) \overset{\text{def}}{=} \frac{\mathrm{p}(X = x, Y = y)}{\mathrm{p}(X = x)}$$

and the conditional expectation of Y given $X = x$ is:

$$\mathrm{E}[Y \mid X = x] \overset{\text{def}}{=} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y \mid X = x)$$

Exm

Example 3.7 (Conditional expectation, one discrete variable and one continuous variable)**.** **X discrete, Y continuous.** The `plasma` dataset in the `dobson` package (Dobson and Barnett (2018), Table 6.25) records plasma inorganic phosphate levels (mg/dL) one hour after a glucose tolerance test, for hyperinsulinemic obese (`H-O`) and control (`C`) participants. Let X be group and Y be phosphate level.

```
library(dobson)
library(dplyr)
data(plasma)
plasma_subset <- plasma |> filter(Group %in% c("H-O", "C"))
plasma_summary <- plasma_subset |>
  summarise(n = n(), mean = mean(phosphate), sd = sd(phosphate), .by = Group)
n_ho <- plasma_summary$n[plasma_summary$Group == "H-O"]
n_c <- plasma_summary$n[plasma_summary$Group == "C"]
mean_ho <- plasma_summary$mean[plasma_summary$Group == "H-O"]
sd_ho <- plasma_summary$sd[plasma_summary$Group == "H-O"]
mean_c <- plasma_summary$mean[plasma_summary$Group == "C"]
sd_c <- plasma_summary$sd[plasma_summary$Group == "C"]
pander::pander(plasma_summary)
```

Group	n	mean	sd
H-O	11	3.945	0.7776
C	12	2.783	0.4086

Modeling phosphate as approximately normal within each group, with parameters set equal to each group's sample mean and SD, and $\mathrm{P}(X = \text{H-O}) = 11/23$:

$$\mathrm{p}(Y = y \mid X = \text{H-O}) \approx \mathrm{N}(3.95,\, 0.78^2)$$
$$\mathrm{p}(Y = y \mid X = \text{C}) \approx \mathrm{N}(2.78,\, 0.41^2)$$

By Definition 3.9, since the mean of a normal random variable is its own location parameter:

$$\mathrm{E}[Y \mid X = \text{H-O}] = 3.95 \text{ mg/dL}, \qquad \mathrm{E}[Y \mid X = \text{C}] = 2.78 \text{ mg/dL}$$

X continuous, Y discrete. The `senility` dataset in the `dobson` package (Dobson and Barnett (2018), Table 7.8) records, for 54 elderly people, a WAIS (Wechsler Adult Intelligence Scale) score and whether symptoms of senility were present. Let X be WAIS score and Y indicate senility symptoms.

```
data(senility)
senility_fit <- glm(s ~ x, data = senility, family = binomial)
b0 <- coef(senility_fit)[["(Intercept)"]]
b1 <- coef(senility_fit)[["x"]]
pander::pander(coef(senility_fit))
```

(Intercept)	x
2.404	-0.3235

Modeling $\mathrm{P}(Y = 1 \mid X = x)$ with the fitted logistic regression:

$$\operatorname{logit} \mathrm{P}(Y = 1 \mid X = x) = 2.4040 - 0.3235\, x$$

By Definition 3.9, since $Y \mid X = x$ is Bernoulli:

$$\begin{aligned} \mathrm{E}[Y \mid X = x] &= 0 \cdot \mathrm{P}(Y = 0 \mid X = x) + 1 \cdot \mathrm{P}(Y = 1 \mid X = x) \\ &= \mathrm{P}(Y = 1 \mid X = x) \\ &= \frac{1}{1 + \mathrm{e}^{-(2.4040 - 0.3235\, x)}} \end{aligned}$$

```
senility_p10 <- predict(senility_fit,
                        newdata = data.frame(x = 10), type = "response")
senility_p15 <- predict(senility_fit,
                        newdata = data.frame(x = 15), type = "response")
```

At $x = 10$: $\mathrm{E}[Y \mid X = 10] = 0.3$. At $x = 15$: $\mathrm{E}[Y \mid X = 15] = 0.08$ — a lower WAIS score is associated with higher predicted probability of senility symptoms.

Definition 3.10 (Conditional expectation function)**.** The **conditional expectation function** $\mathrm{E}[Y \mid X]$ is the function (and hence random variable) of X obtained by evaluating $\mathrm{E}[Y \mid X = x]$ at X; specifically, $\mathrm{E}[Y \mid X] = g(X)$ where $g(x) \stackrel{\text{def}}{=} \mathrm{E}[Y \mid X = x]$.

Exm

Example 3.8 (The conditional expectation function of the birthweight model)**.** Continuing Example 3.5 and Example 3.6, (X, Y) (gestational age, birthweight) is modeled as bivariate normal. The general form of the conditional mean derived in Example 3.5, evaluated at an arbitrary x instead of just $x = 40$, gives the conditional expectation function directly:

$$\begin{aligned} g(x) &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) \\ &= -1484.9846 + 115.5283\,x \end{aligned}$$

which is exactly the fitted regression line (the general algebraic identity $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) = \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \rho\frac{\sigma_Y}{\sigma_X}x$ is what makes the bivariate-normal conditional mean linear in x). As a check, evaluating at $x = 40$ recovers Example 3.6's result:

$$g(40) = -1484.9846 + 115.5283 \times 40 = 3136.15 \text{ g}$$

Exercise 3.1 (Expectation of a sum, given a joint PMF)**.** Let (X, Y) be discrete with joint probability mass function:

	$Y = 0$	$Y = 1$
$X = 0$	0.2	0.3
$X = 1$	0.1	0.4

Compute $\mathrm{E}[X + Y]$.

Solution

Solution. Treating (X, Y) as a single discrete random object taking one of the four values $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, LOTUS (Theorem 3.5) applied to $g(x, y) = x + y$ gives directly:

$$\begin{aligned} \mathrm{E}[X+Y] &= \sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} (x+y)\ \mathrm{P}(X=x,\,Y=y) \\ &= (0{+}0)(0.2) + (0{+}1)(0.3) + (1{+}0)(0.1) + (1{+}1)(0.4) \\ &= 0 + 0.3 + 0.1 + 0.8 \\ &= 1.2 \end{aligned}$$

As a check: $\mathrm{E}[X] = 0(0.5) + 1(0.5) = 0.5$ and $\mathrm{E}[Y] = 0(0.3) + 1(0.7) = 0.7$, so $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 1.2$.

```r
x_labs <- c("X=0", "X=0", "X=1", "X=1")
y_labs <- c("Y=0", "Y=1", "Y=0", "Y=1")
probs <- c(0.2, 0.3, 0.1, 0.4)

plotly::plot_ly(
  x = ~y_labs, y = ~probs, color = ~x_labs,
  colors = c("steelblue", "tomato"),
  type = "bar"
) |>
  plotly::layout(
    barmode = "group",
    xaxis = list(title = "Y"),
    yaxis = list(title = "P(X = x, Y = y)", range = c(0, 0.5)),
    legend = list(title = list(text = "X value"))
  )
```

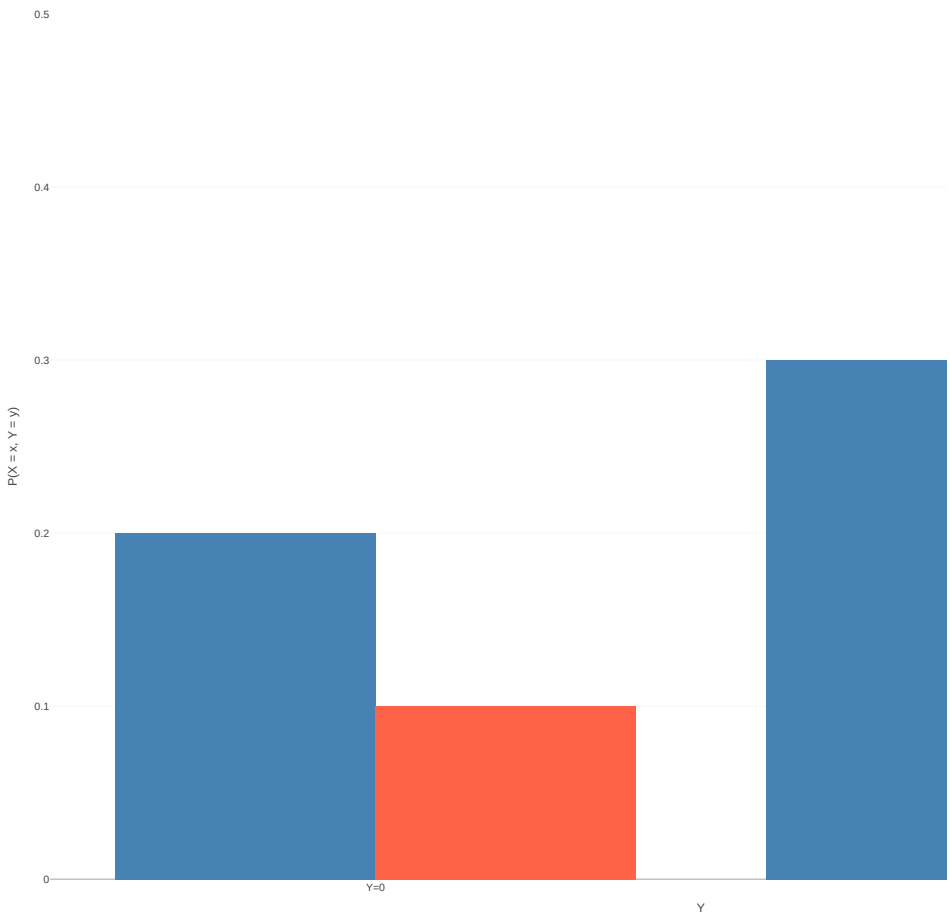


Figure 4: Joint probability mass function $\mathrm{P}(X = x, Y = y)$. Marginal totals: $\mathrm{P}(X = 0) = 0.5$, $\mathrm{P}(X = 1) = 0.5$, $\mathrm{P}(Y = 0) = 0.3$, $\mathrm{P}(Y = 1) = 0.7$.

Note that X and Y are **not** independent here: $\mathrm{P}(X=0, Y=0) = 0.2 \neq 0.15 = \mathrm{P}(X=0)\ \mathrm{P}(Y=0)$. LOTUS applies regardless, since it requires only the *actual* joint mass function, not independence.

There are only four (x, y) pairs here, so summing them in any order — row by row, column by column, or any other listing — gives the same total by ordinary commutativity and associativity of addition; no result about interchanging summation order is needed. Fubini–Tonelli, covered next, is what's needed once the support is countably infinite instead — see Exercise 3.2.

3.4 Fubini–Tonelli for expectations

The Riemann version of Fubini's theorem[4], stated in the math-prereqs chapter[5], lets us switch the order of integration for continuous integrands on simple regions. For expectations against probability measures we use its measure-theoretic form[6], which holds on any -finite measure space. The -finiteness condition is automatic for probability measures (every probability measure is finite, hence -finite), so Fubini–Tonelli[7] yields Corollary 3.1 directly.

Corollary 3.1 (Joint-distribution form (without independence; corollary of Fubini–Tonelli))**.** *Let (X, Y) be jointly distributed random variables whose joint distribution has a density $f_{X,Y}$ with respect to a product of -finite reference measures $\mu_X \otimes \mu_Y$ on $\mathcal{R}(X) \times \mathcal{R}(Y)$, and let $h : \mathcal{R}(X) \times \mathcal{R}(Y) \to \mathbb{R}$ be measurable. If either*

(a) $h(X, Y) \geq 0$ almost surely, or

(b) $\mathrm{E}[|h(X, Y)|] < \infty$,

then the expectation of $h(X, Y)$ can be written as an iterated integral against $f_{X,Y}$, with the order of integration exchangeable:

$$\begin{aligned}\mathrm{E}[h(X,Y)] &= \int_{\mathcal{R}(X)} \left(\int_{\mathcal{R}(Y)} h(x,y)\, f_{X,Y}(x,y)\, d\mu_Y(y) \right) d\mu_X(x) \\ &= \int_{\mathcal{R}(Y)} \left(\int_{\mathcal{R}(X)} h(x,y)\, f_{X,Y}(x,y)\, d\mu_X(x) \right) d\mu_Y(y).\end{aligned}$$

The choice of reference measures covers three cases:

- ***Both continuous:*** *$\mu_X = \mu_Y =$ Lebesgue measure; $f_{X,Y}$ is the joint probability density function (PDF), and $\int g(x)\, d\mu_X(x) = \int g(x)\, dx$.*
- ***Both discrete:*** *$\mu_X = \mu_Y =$ counting measure; $f_{X,Y}(x, y) = \mathrm{P}(X = x,\, Y = y)$ is the joint probability mass function (PMF), and $\int g(x)\, d\mu_X(x) = \sum_{x \in \mathcal{R}(X)} g(x)$.*
- ***Mixed*** *(one continuous, one discrete): one reference measure is Lebesgue and the other is counting; $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\ \mathrm{P}(Y = y)$ (or $\mathrm{P}(X = x \mid Y = y)\, f_Y(y)$ if X is discrete and Y continuous), and the iterated integrals combine an ordinary integral with a sum. The conditional densities/PMFs here are defined the same way as in Definition 3.9, just conditioning on Y instead of X.*

Proof

Proof. Apply Fubini–Tonelli[a] with $\mu_1 = \mu_X$ and $\mu_2 = \mu_Y$ to the integrand $h(x, y)\, f_{X,Y}(x, y)$ on $\mathcal{R}(X) \times \mathcal{R}(Y)$. Lebesgue measure and counting measure on a countable set are each -finite, so $\mu_X \otimes \mu_Y$ is -finite in all three cases. The relevant condition is (a) when $h \geq 0$ and (b) when $\mathrm{E}[|h(X, Y)|] < \infty$. Independence is not required. When X and Y are independent, $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$ (or $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\ \mathrm{P}(Y = y)$ in the discrete case), and the iterated integrals factor into separate integrals over the marginals. □

[a] math-prereqs.qmd#thm-fubini-tonelli

[4] math-prereqs.qmd#thm-fubini
[5] math-prereqs.qmd
[6] math-prereqs.qmd#thm-fubini-tonelli
[7] math-prereqs.qmd#thm-fubini-tonelli

Exm

Example 3.9 (Expectation of a product of independent variables)**.** Let $X \sim \text{Uniform}(0,1)$ and $Y \sim \text{Uniform}(0,2)$, independently distributed. Compute $\mathrm{E}[XY]$.
We apply Corollary 3.1 (both-continuous case) with $h(x,y) = xy$. Since X and Y are independent with densities $f_X(x) = 1$ on $[0,1]$ and $f_Y(y) = \frac{1}{2}$ on $[0,2]$, the joint density factors as $f_{X,Y}(x,y) = f_X(x)\, f_Y(y) = \frac{1}{2}$, and $\mu_X = \mu_Y =$ Lebesgue measure:

$$\begin{aligned}
\mathrm{E}[XY] &= \int_0^1 \left(\int_0^2 xy \cdot \tfrac{1}{2}\, dy \right) dx \\
&= \int_0^1 x \left(\frac{1}{2} \int_0^2 y\, dy \right) dx \\
&= \int_0^1 x \cdot \frac{1}{2} \cdot \left[\frac{y^2}{2} \right]_0^2 dx \\
&= \int_0^1 x \cdot \frac{1}{2} \cdot 2\, dx \\
&= \int_0^1 x\, dx \\
&= \frac{1}{2}
\end{aligned}$$

As a check: $\mathrm{E}[X] = \frac{1}{2}$, $\mathrm{E}[Y] = 1$, and $\mathrm{E}[X]\,\mathrm{E}[Y] = \frac{1}{2}$.

Exm

Example 3.10 (When independence fails: a counterexample)**.** Correctly applying Corollary 3.1 requires the *actual* joint density $f_{X,Y}$ — not the product of marginals $f_X(x)\, f_Y(y)$, which is valid only when X and Y are independent. Using the wrong joint density gives the wrong answer.
Let $X \sim \text{Uniform}(0,1)$ and set $Y = X$ (so X and Y are perfectly correlated and **not** independent).
True expectation:

$$\mathrm{E}[XY] = \mathrm{E}[X \cdot X] = \mathrm{E}[X^2] = \int_0^1 x^2\, dx = \frac{1}{3}$$

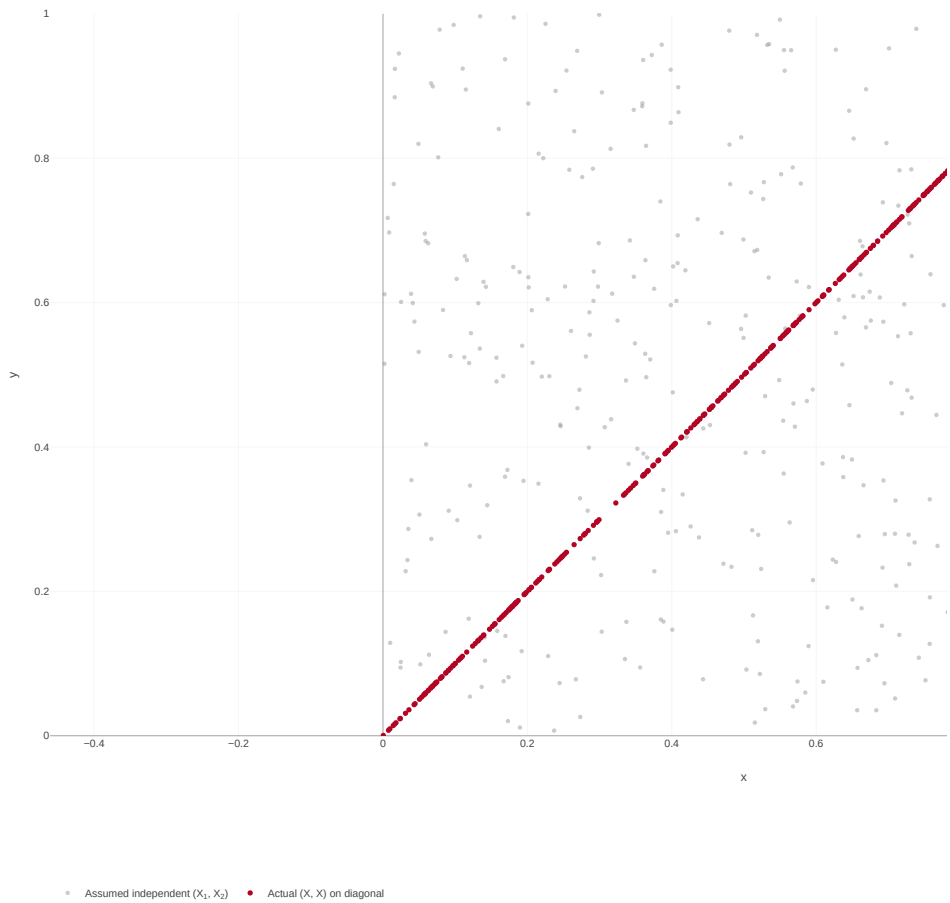
Erroneously applying the product-measure formula:
Note that Fubini–Tonelli's own conditions still hold here ($h(x,y) = xy$ is nonnegative and integrable), so the error is not a failure of Fubini–Tonelli[a]. Rather, the error is using the *wrong measure*: the joint distribution of (X,X) is concentrated on the diagonal $\{(x,x) : x \in [0,1]\} \subset [0,1]^2$, which has Lebesgue measure zero in $\mathbb{R}^2$. The joint distribution is therefore **not** absolutely continuous with respect to two-dimensional Lebesgue measure, so **no joint density $f_{X,Y}$ on $[0,1]^2$ exists**, which is the reference density Corollary 3.1 requires.
The following calculation is what someone would *erroneously* write if they assumed independence and used $f_X(x)\, f_Y(y)$ as a "joint density" — a function that does not in fact correspond to the joint distribution of (X,X). The marginals $X \sim \text{Uniform}(0,1)$ and $Y \sim \text{Uniform}(0,1)$ do have densities $f_X = f_Y = 1$, but the *product* $f_X(x)\, f_Y(y) = 1$ on $[0,1]^2$ is the density of an *independent* pair, not of (X,X):

$$\begin{aligned}\int_0^1 \int_0^1 xy \cdot f_X(x) \cdot f_Y(y)\, dy\, dx &= \int_0^1 \int_0^1 xy\, dy\, dx \\ &= \int_0^1 x \left(\int_0^1 y\, dy \right) dx \\ &= \int_0^1 x \cdot \frac{1}{2}\, dx \\ &= \frac{1}{4}\end{aligned}$$

This calculation recovers $\mathrm{E}[XY]$ for *independent* uniforms ($\frac{1}{4}$), not $\mathrm{E}[XX]$ for the perfectly correlated pair ($\frac{1}{3}$). The lesson is that Corollary 3.1 requires the *actual* joint density $f_{X,Y}$. For independent (X, Y), this factors as $f_X(x)\, f_Y(y)$; for dependent (X, Y), $f_{X,Y}$ need not factor — and for (X, X), no joint density on $\mathbb{R}^2$ exists at all, so Corollary 3.1 simply does not apply.

```r
set.seed(204)
n <- 400
x_dep <- runif(n)
y_dep <- x_dep
x_ind <- runif(n)
y_ind <- runif(n)

plotly::plot_ly() |>
  plotly::add_trace(
    type = "scatter", mode = "markers",
    x = x_ind, y = y_ind,
    name = "Assumed independent (X<sub>1</sub>, X<sub>2</sub>)",
    marker = list(size = 5, color = "#999999", opacity = 0.5)
  ) |>
  plotly::add_trace(
    type = "scatter", mode = "markers",
    x = x_dep, y = y_dep,
    name = "Actual (X, X) on diagonal",
    marker = list(size = 6, color = "#b40426")
  ) |>
  plotly::layout(
    xaxis = list(title = "x", range = c(0, 1), scaleanchor = "y"),
    yaxis = list(title = "y", range = c(0, 1)),
    legend = list(orientation = "h", y = -0.2)
  )
```



[a] math-prereqs.qmd#thm-fubini-tonelli

Exm

Example 3.11 (Both-continuous case: joint PDF on a non-rectangular support)**.** Let (X, Y) have joint density $f_{X,Y}(x, y) = 2$ for $0 \le x \le y \le 1$ (and 0 otherwise). Compute $\mathrm{E}[X + Y]$. By Corollary 3.1:

$$\begin{aligned}
\mathrm{E}[X+Y] &= \int_0^1 \int_0^y (x+y) \cdot 2 \, dx \, dy \\
&= 2 \int_0^1 \left[\frac{x^2}{2} + xy \right]_{x=0}^{x=y} dy \\
&= 2 \int_0^1 \left(\frac{y^2}{2} + y^2 \right) dy \\
&= 2 \int_0^1 \frac{3y^2}{2} \, dy \\
&= 3 \int_0^1 y^2 \, dy \\
&= 3 \cdot \frac{1}{3} \\
&= 1
\end{aligned}$$

```r
n_grid <- 51
x_seq <- seq(0, 1, length.out = n_grid)
y_seq <- seq(0, 1, length.out = n_grid)

z_mat <- outer(x_seq, y_seq, function(x, y) {
  z <- rep(2, length(x))
  z[x > y] <- NA
  z
})

plotly::plot_ly(x = ~x_seq, y = ~y_seq, z = ~t(z_mat)) |>
  plotly::add_surface(showscale = FALSE) |>
  plotly::layout(scene = list(
    xaxis = list(title = "x"),
    yaxis = list(title = "y"),
    zaxis = list(title = "f(x, y)", range = c(0, 2.5)),
    camera = list(eye = list(x = 1.6, y = -1.6, z = 0.8))
  ))
```

Exercise 3.2 (Both-discrete case, infinite support: joint PMF)**.** Let X and Y be independent, each Geometric on $\mathcal{R}(X) = \mathcal{R}(Y) = \{0, 1, 2, \dots\}$ with $\mathrm{P}(X = x) = (1-p)\,p^x$ for a fixed $p \in (0, 1)$ (X counts the number of failures before the first success in a sequence of independent trials with success probability $1 - p$; likewise for Y). Unlike Exercise 3.1, the support here is countably infinite. The joint PMF is $\mathrm{P}(X = x,\, Y = y) = (1-p)^2\,p^{x+y}$.
Compute $\mathrm{E}[X + Y]$.

Solution

Solution. Compute $\mathrm{E}[X + Y]$ using Corollary 3.1 with $\mu_X = \mu_Y =$ counting measure and $h(x, y) = x + y$. Since $h(x, y) = x + y \geq 0$ for every (x, y) in this support, condition (a) holds, so Corollary 3.1 (via Tonelli's theorem) guarantees the order of this now-infinite double sum is exchangeable — unlike the finite case, elementary algebra alone could not establish this.
By Corollary 3.1 (both-discrete case), summing over y first for each fixed x:

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(x+y)\ \mathrm{P}(X=x,\,Y=y)\\
&= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(x+y)(1-p)^2p^{x+y}\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\sum_{y=0}^{\infty}(x+y)\,p^y\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\left(x\sum_{y=0}^{\infty}p^y+\sum_{y=0}^{\infty}y\,p^y\right)\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\left(\frac{x}{1-p}+\frac{p}{(1-p)^2}\right)\\
&= \sum_{x=0}^{\infty}p^x[x(1-p)+p]\\
&= (1-p)\sum_{x=0}^{\infty}x\,p^x+p\sum_{x=0}^{\infty}p^x\\
&= (1-p)\cdot\frac{p}{(1-p)^2}+p\cdot\frac{1}{1-p}\\
&= \frac{p}{1-p}+\frac{p}{1-p}\\
&= \frac{2p}{1-p}
\end{aligned}$$

using the standard geometric-series facts $\sum_{y=0}^{\infty}p^y = \frac{1}{1-p}$ and $\sum_{y=0}^{\infty}y\,p^y = \frac{p}{(1-p)^2}$ (e.g. Casella and Berger (2002)).
As a check, $\mathrm{E}[X] = \mathrm{E}[Y] = \frac{p}{1-p}$ (the mean of this Geometric distribution; Casella and Berger (2002)), so $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = \frac{2p}{1-p}$ by linearity, matching.

```r
p <- 0.4
exact_sum <- 2 * p / (1 - p)

n_trunc <- 500
support <- 0:n_trunc
joint_pmf_fn <- function(x, y) (1 - p)^2 * p^(x + y)
joint_probs <- outer(support, support, joint_pmf_fn)
h_vals <- outer(support, support, function(x, y) x + y)

sum_y_first <- sum(rowSums(h_vals * joint_probs))
sum_x_first <- sum(colSums(h_vals * joint_probs))

pander::pander(data.frame(
  quantity = c("exact (closed form)", "sum over y first, then x",
               "sum over x first, then y"),
  value = round(c(exact_sum, sum_y_first, sum_x_first), 6)
))
```

quantity	value
exact (closed form)	1.333
sum over y first, then x	1.333
sum over x first, then y	1.333

With $p = 0.4$, summing y first then x, and summing x first then y, agree with each other and with the closed form $\frac{2p}{1-p} = 1.3333$ up to truncation error — exactly Tonelli's guarantee.

i Tonelli's nonnegativity condition, and what happens without it

The calculation in Exercise 3.2 only needed condition (a), $h(X, Y) \geq 0$, because $h(x, y) = x + y$ is nonnegative on this support. For a **signed** h, interchanging an infinite double sum is not automatically valid — Corollary 3.1's condition (b), $\mathrm{E}[|h(X, Y)|] < \infty$, is what licenses it in that case. Without either condition, the two orders can genuinely disagree. A standard example (a signed array, not a probability distribution; see e.g. Rudin (1976) for the general theory of rearranging series): let $a_{m,n} = 1$ if $m = n$, $a_{m,n} = -1$ if $m = n + 1$, and $a_{m,n} = 0$ otherwise, for $m, n = 0, 1, 2, \ldots$. Summing each row m first: row 0 has only the term $a_{0,0} = 1$ (there is no valid $n = -1$), so its row sum is 1; every row $m \geq 1$ has $a_{m,m} = 1$ and $a_{m,m-1} = -1$, so its row sum is 0. Summing the rows then gives $1 + 0 + 0 + \cdots = 1$. Summing each column n first: every column $n \geq 0$ has $a_{n,n} = 1$ and $a_{n+1,n} = -1$, so its column sum is always 0, and summing the columns then gives $0 + 0 + \cdots = 0$. The two orders give 1 and 0: genuinely different answers, confirming that a condition like (a) or (b) really is needed once the terms are no longer all nonnegative.

Exercise 3.3 (Mixed case: one continuous variable, one discrete variable)**.** Let $Y \sim$ Bernoulli(0.6) and, given $Y = y$, let $X \mid Y = y \sim$ Uniform$(0,\, y + 1)$.
Compute $\mathrm{E}[X]$.

Solution

Solution. Compute $\mathrm{E}[X]$ using Corollary 3.1 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x, y) = x$.
The joint density w.r.t. Lebesgue × counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\ \mathrm{P}(Y = y)$:

$$f_{X,Y}(x,\,0) = 1 \cdot 0.4 = 0.4 \qquad \text{for } x \in [0,1];$$
$$f_{X,Y}(x,\,1) = \tfrac{1}{2} \cdot 0.6 = 0.3 \qquad \text{for } x \in [0,2].$$

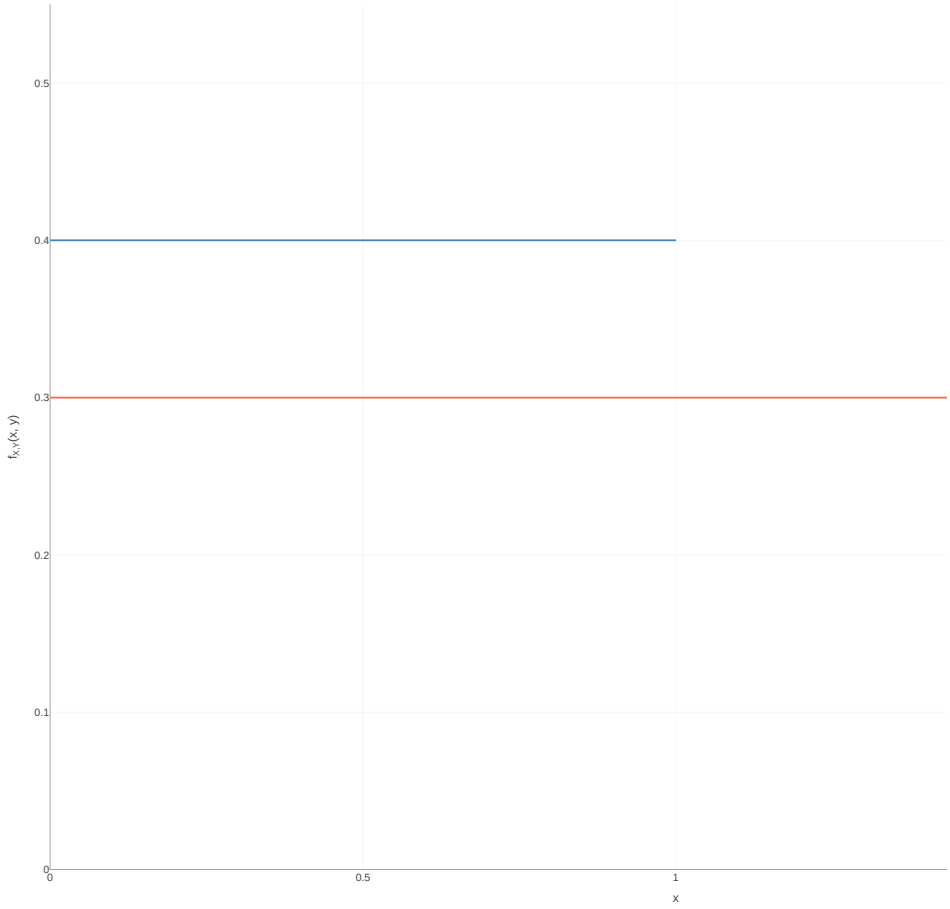
By Corollary 3.1 (mixed case):

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{y \in \{0,1\}} \int_0^{y+1} x\, f_{X,Y}(x,\,y)\, dx \\
&= \int_0^1 x \cdot 0.4\, dx + \int_0^2 x \cdot 0.3\, dx \\
&= 0.4 \cdot \frac{1}{2} + 0.3 \cdot 2 \\
&= 0.2 + 0.6 = 0.8
\end{aligned}$$

As a check using the law of total expectation: $\mathrm{E}[X \mid Y = 0] = \frac{1}{2}$ and $\mathrm{E}[X \mid Y = 1] = 1$, so $\mathrm{E}[X] = \frac{1}{2}(0.4) + 1(0.6) = 0.2 + 0.6 = 0.8$.

```r
x_fine <- seq(0, 2, by = 0.005)
df <- data.frame(
  x = c(x_fine[x_fine <= 1], x_fine),
  density = c(rep(0.4, sum(x_fine <= 1)), rep(0.3, length(x_fine))),
  label = c(
    rep("Y = 0  (P = 0.4)", sum(x_fine <= 1)),
    rep("Y = 1  (P = 0.6)", length(x_fine))
  )
)

plotly::plot_ly(
  df, x = ~x, y = ~density, color = ~label,
  colors = c("steelblue", "tomato")
) |>
  plotly::add_lines() |>
  plotly::layout(
    xaxis = list(title = "x"),
    yaxis = list(title = "f<sub>X,Y</sub>(x, y)", range = c(0, 0.55)),
    legend = list(title = list(text = "Y value"))
  )
```



Y takes only finitely many values here (two), so $\sum_{y\in\{0,1\}} \int_0^{y+1} x\, f_{X,Y}(x,\, y)\, dx$ is just linearity of the integral applied to a two-term sum — $\int g + \int k = \int (g+k)$ — not a genuine interchange of summation and integration order. Exercise 3.4 gives Y a countably infinite range instead, so the sum-of-integrals is genuinely infinite and Corollary 3.1's guarantee is load-bearing.

Exercise 3.4 (Mixed case, infinite discrete support)**.** Let Y be Geometric on $\{0, 1, 2, \dots\}$ with $\mathrm{P}(Y = y) = (1-q)\, q^y$ for a fixed $q \in (0, 1)$ and, given $Y = y$, let $X \mid Y = y \sim \mathrm{Uniform}(0,\, y+1)$. Unlike Exercise 3.3, Y's range here is countably infinite.
Compute $\mathrm{E}[X]$.

Solution

Solution. Compute $\mathrm{E}[X]$ using Corollary 3.1 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x, y) = x$.
The joint density w.r.t. Lebesgue × counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\ \mathrm{P}(Y = y) = \frac{(1-q)\, q^y}{y+1}$ for $x \in [0, y+1]$.
Since $h(x, y) = x \geq 0$ on this support, condition (a) holds, so Corollary 3.1 (via Tonelli's theorem) guarantees the now-infinite sum-of-integrals expression is valid — unlike in Exercise 3.3, this Fubini–Tonelli justification is required because the sum is infinite rather than finite.
By Corollary 3.1 (mixed case):

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{y=0}^{\infty} \int_0^{y+1} x\, f_{X,Y}(x,\, y)\, dx \\
&= \sum_{y=0}^{\infty} \frac{(1-q)\, q^y}{y+1} \int_0^{y+1} x\, dx \\
&= \sum_{y=0}^{\infty} \frac{(1-q)\, q^y}{y+1} \cdot \frac{(y+1)^2}{2} \\
&= \frac{1-q}{2} \sum_{y=0}^{\infty} (y+1)\, q^y \\
&= \frac{1-q}{2} \left(\sum_{y=0}^{\infty} y\, q^y + \sum_{y=0}^{\infty} q^y \right) \\
&= \frac{1-q}{2} \left(\frac{q}{(1-q)^2} + \frac{1}{1-q} \right) \\
&= \frac{1-q}{2} \cdot \frac{q + (1-q)}{(1-q)^2} \\
&= \frac{1-q}{2} \cdot \frac{1}{(1-q)^2} \\
&= \frac{1}{2(1-q)}
\end{aligned}$$

using the same geometric-series facts as Exercise 3.2 (e.g. Casella and Berger (2002)).
As a check using the law of total expectation: $\mathrm{E}[X \mid Y = y] = \frac{y+1}{2}$ (the mean of $\mathrm{Uniform}(0, y+1)$) and $\mathrm{E}[Y] = \frac{q}{1-q}$ (the mean of this Geometric distribution; Casella and Berger (2002)), so:

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}[\mathrm{E}[X \mid Y]] \\
&= \mathrm{E}\left[\frac{Y+1}{2}\right] \\
&= \frac{\mathrm{E}[Y]+1}{2} \\
&= \frac{1}{2}\left(\frac{q}{1-q}+1\right) \\
&= \frac{1}{2}\cdot\frac{q+(1-q)}{1-q} \\
&= \frac{1}{2(1-q)}
\end{aligned}$$

matching.

```
q <- 0.5
exact_ex <- 1 / (2 * (1 - q))

n_trunc <- 2000
y_vals <- 0:n_trunc
py_vals <- (1 - q) * q^y_vals
cond_mean <- (y_vals + 1) / 2
trunc_ex <- sum(cond_mean * py_vals)

pander::pander(data.frame(
  quantity = c("exact (closed form)", "truncated sum-of-integrals"),
  value = round(c(exact_ex, trunc_ex), 6)
))
```

quantity	value
exact (closed form)	1
truncated sum-of-integrals	1

With $q = 0.5$, the truncated sum-of-integrals matches the closed form $\frac{1}{2(1-q)} = 1$.

Theorem 3.6 (Law of iterated expectations)**.** *For any two random variables X and Y:*

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]$$

Alternate names for this identity include: the ***tower rule****, the* ***tower property****, the* ***law of total expectation****, and the* ***smoothing theorem****.*

Proof

Proof. **Discrete case.** When X and Y are discrete, applying Definition 3.5 to $\mathrm{E}[\mathrm{E}[Y \mid X]]$ and then the law of total probability (Theorem 1.4) applied to the countable partition $\{X = x : x \in \mathcal{R}(X)\}$:

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X]] &= \sum_{x \in \mathcal{R}(X)} \mathrm{E}[Y \mid X = x] \cdot \mathrm{P}(X = x) && \text{(definition of expectation of } \mathrm{E}[Y \mid X]) \\
&= \sum_{x \in \mathcal{R}(X)} \left(\sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y \mid X = x) \right) \cdot \mathrm{P}(X = x) && \text{(definition of conditional expectation } \mathrm{E}[Y \mid X = x]) \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{x \in \mathcal{R}(X)} \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{P}(X = x) && \text{(exchange order of summation)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y) && \text{(law of total probability over } X) \\
&= \mathrm{E}[Y] && \text{(definition of expectation of } Y)
\end{aligned}
$$

Continuous case. When X and Y are continuous, applying Definition 3.5 to $\mathrm{E}[\mathrm{E}[Y \mid X]]$ and then using Definition 3.8 for $\mathrm{E}[Y \mid X = x]$:

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X]] &= \int_{x \in \mathcal{R}(X)} \mathrm{E}[Y \mid X = x] \cdot \mathrm{p}(X = x)\, dx \\
&= \int_{x \in \mathcal{R}(X)} \left(\int_{y \in \mathcal{R}(Y)} y \cdot \mathrm{p}(Y = y \mid X = x)\, dy \right) \cdot \mathrm{p}(X = x)\, dx \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \left(\int_{x \in \mathcal{R}(X)} \mathrm{p}(Y = y \mid X = x) \cdot \mathrm{p}(X = x)\, dx \right) dy \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \mathrm{p}(Y = y)\, dy \\
&= \mathrm{E}[Y]
\end{aligned}
$$

where the third equality exchanges the order of integration by condition (b) of Fubini–Tonelli[a] (the absolute-integrability case, **Fubini's theorem**); this theorem requires $\mathrm{E}[|Y|] < \infty$, which is implicit in $\mathrm{E}[Y]$ being defined, and the fourth equality uses $\int_x \mathrm{p}(Y = y \mid X = x) \cdot \mathrm{p}(X = x)\, dx = \int_x \mathrm{p}(X = x, Y = y)\, dx = \mathrm{p}(Y = y)$ (marginalization of the joint density). □

[a] math-prereqs.qmd#thm-fubini-tonelli

Theorem 3.7 (Conditional law of iterated expectations)**.** *For random variables X, Y, and Z:*

$$\mathrm{E}[Y \mid Z] = \mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z]$$

This identity is the tower rule applied conditionally on Z.

Proof

Proof. For each fixed value z with positive probability or density:
Discrete case. Conditioning on $Z = z$, and applying the law of total probability to the partition $\{X = x : x \in \mathcal{R}(X)\}$ under the conditional distribution given $Z = z$:

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z = z] &= \sum_{x \in \mathcal{R}(X)} \mathrm{E}[Y \mid X = x, Z = z] \cdot \mathrm{P}(X = x \mid Z = z) && \text{(definition of expectation g} \\
&= \sum_{x \in \mathcal{R}(X)} \left(\sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y \mid X = x, Z = z) \right) \cdot \mathrm{P}(X = x \mid Z = z) && \text{(definition of conditional e} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{x \in \mathcal{R}(X)} \mathrm{P}(Y = y \mid X = x, Z = z) \cdot \mathrm{P}(X = x \mid Z = z) && \text{(exchange order of summa} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y \mid Z = z) && \text{(law of total probability gi} \\
&= \mathrm{E}[Y \mid Z = z] && \text{(definition of conditional e}
\end{aligned}
$$

Continuous case. Conditioning on $Z = z$, and integrating over X under the conditional density $\mathrm{p}(X = x \mid Z = z)$:

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z = z] &= \int_{x \in \mathcal{R}(X)} \mathrm{E}[Y \mid X = x, Z = z] \cdot \mathrm{p}(X = x \mid Z = z)\, dx \\
&= \mathrm{E}[Y \mid Z = z]
\end{aligned}
$$

Therefore, as random variables of Z, $\mathrm{E}[Y \mid Z] = \mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z]$. □

Exm

Example 3.12 (Marginal expectation from conditional expectations)**.** Suppose X is a binary random variable indicating treatment assignment ($X = 1$ treated, $X = 0$ control), with $\mathrm{P}(X = 1) = 0.5$, and suppose the outcome Y has conditional expectations:

$$
\mathrm{E}[Y \mid X = 1] = 10, \quad \mathrm{E}[Y \mid X = 0] = 6
$$

By the law of iterated expectations (Theorem 3.6):

$$
\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}[Y \mid X]] \\
&= \mathrm{E}[Y \mid X = 1] \cdot \mathrm{P}(X = 1) + \mathrm{E}[Y \mid X = 0] \cdot \mathrm{P}(X = 0) \\
&= 10 \cdot 0.5 + 6 \cdot 0.5 \\
&= 5 + 3 \\
&= 8
\end{aligned}
$$

Definition 3.11 (Expectation of a random matrix)**.** For a random matrix $\mathbf{A}$ of size $m \times n$ with (i, j)-th element A_{ij}, the **expectation** $\mathrm{E}\,\mathbf{A}$ is the $m \times n$ matrix whose (i, j)-th element is $\mathrm{E}[A_{ij}]$:

$$
\mathrm{E}\,\mathbf{A} \overset{\text{def}}{=} \begin{pmatrix} \mathrm{E}[A_{11}] & \mathrm{E}[A_{12}] & \cdots & \mathrm{E}[A_{1n}] \\ \mathrm{E}[A_{21}] & \mathrm{E}[A_{22}] & \cdots & \mathrm{E}[A_{2n}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[A_{m1}] & \mathrm{E}[A_{m2}] & \cdots & \mathrm{E}[A_{mn}] \end{pmatrix}
$$

In other words, expectation is applied **element-wise** to a random matrix.

3.5 Deviation, error, and noise

Definition 3.12 (Deviation)**.** A **deviation** is the difference between a value and a reference value. For any quantity z and reference value r:

$$z - r$$

In probability and statistics, "deviation" often means deviation from a population mean. For a random variable Y:

$$Y - \mathrm{E}[Y]$$

See: Wikipedia: Deviation (statistics)[a]

[a] https://en.wikipedia.org/wiki/Deviation_(statistics)

Definition 3.13 (Deviation from a population or subpopulation mean)**.** In probabilistic models, we call this quantity a **deviation from a mean**. It is often also called an **error** or **noise term** in other sources. For the random variable Y, define the deviation from its mean as:

$$e(Y) \stackrel{\text{def}}{=} Y - \mathrm{E}[Y]$$

For a realized observation y:

$$e(y) \stackrel{\text{def}}{=} y - \mathrm{E}[Y]$$

In regression settings, the reference mean is often conditional on covariates: $e(y_i) \stackrel{\text{def}}{=} y_i - \mathrm{E}[Y_i \mid X_i]$.
In this course, we prefer "deviation" for this mean-deviation quantity. The terms "error" and "noise" are common aliases. We use "residual" (defined in the Linear regression chapter[a]) for deviations from fitted values. For notation in this course, we use $e(\cdot)$ for these model/data deviations, and reserve $\varepsilon(\cdot)$ for estimator-to-estimand deviations (see Estimation[b]).
See:

- Wikipedia: Errors and residuals[c]
- Wikipedia: Deviation (statistics)[d]
- Wikipedia: Linear regression — Notation and terminology[e]

[a] Linear-models-overview.qmd#def-resid-fitted
[b] estimation.qmd#def-estimation-error
[c] https://en.wikipedia.org/wiki/Errors_and_residuals
[d] https://en.wikipedia.org/wiki/Deviation_(statistics)
[e] https://en.wikipedia.org/wiki/Linear_regression#Notation_and_terminology

3.6 Variance and related characteristics

Definition 3.14 (Variance)**.** The variance of a random variable X is the expectation of the squared deviation from the mean; that is:

$$\mathrm{Var}(X) \stackrel{\text{def}}{=} \mathrm{E}[[e(X)]^2]$$

Theorem 3.8 (Variance as expected squared deviation from the mean)**.**

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

Proof

Proof. Substituting the definition of $e(X)$ from Definition 3.13 into Definition 3.14:

$$\mathrm{Var}(X) \stackrel{\text{def}}{=} \mathrm{E}[[e(X)]^2] = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

□

Theorem 3.9 (Simplified expression for variance)**.**

$$\operatorname{Var}(X) = \operatorname{E}[X^2] - (\operatorname{E}[X])^2$$

Proof

Proof. By linearity of expectation, we have:

$$\begin{aligned}
\operatorname{Var}(X) &\stackrel{\text{def}}{=} \operatorname{E}[[e(X)]^2] && \text{(definition of variance)}\\
&= \operatorname{E}[(X - \operatorname{E}[X])^2] && \text{(definition of deviation from mean)}\\
&= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + (\operatorname{E}[X])^2\right] && \text{(expand binomial square)}\\
&= \operatorname{E}[X^2] - \operatorname{E}[2X\operatorname{E}[X]] + \operatorname{E}\left[(\operatorname{E}[X])^2\right] && \text{(linearity of expectation)}\\
&= \operatorname{E}[X^2] - 2\operatorname{E}[X]\operatorname{E}[X] + (\operatorname{E}[X])^2 && \text{(constants factor out of expectation)}\\
&= \operatorname{E}[X^2] - (\operatorname{E}[X])^2 && \text{(algebraic simplification)}
\end{aligned}$$

□

Theorem 3.10 (Law of total variance)**.** *For random variables X and Y:*

$$\operatorname{Var}(Y) = \operatorname{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\operatorname{E}[Y \mid X])$$

where $\operatorname{Var}(Y \mid X) \stackrel{def}{=} \operatorname{E}[(Y - \operatorname{E}[Y \mid X])^2 \mid X]$.
Alternate names include: the ***conditional variance formula****,* ***Eve's law****, and the* ***variance decomposition formula****.*

Proof

Proof. Write $Y - \operatorname{E}[Y] = (Y - \operatorname{E}[Y \mid X]) + (\operatorname{E}[Y \mid X] - \operatorname{E}[Y])$. Then:

$$(Y - \operatorname{E}[Y])^2 = (Y - \operatorname{E}[Y \mid X])^2 + (\operatorname{E}[Y \mid X] - \operatorname{E}[Y])^2 + 2(Y - \operatorname{E}[Y \mid X])(\operatorname{E}[Y \mid X] - \operatorname{E}[Y])$$

Taking expectation:

$$\begin{aligned}
\operatorname{Var}(Y) = \operatorname{E}\left[(Y - \operatorname{E}[Y \mid X])^2\right] + \operatorname{E}\left[(\operatorname{E}[Y \mid X] - \operatorname{E}[Y])^2\right]\\
+ 2\operatorname{E}[(Y - \operatorname{E}[Y \mid X])(\operatorname{E}[Y \mid X] - \operatorname{E}[Y])]
\end{aligned}$$

For the cross-term:
Discrete case.

$$\begin{aligned}
\operatorname{E}[(Y - \operatorname{E}[Y \mid X])(\operatorname{E}[Y \mid X] - \operatorname{E}[Y])] &= \sum_{x \in \mathcal{R}(X)} \operatorname{E}[(Y - \operatorname{E}[Y \mid X])(\operatorname{E}[Y \mid X] - \operatorname{E}[Y]) \mid X = x] \cdot \operatorname{P}(X = x)\\
&= \sum_{x \in \mathcal{R}(X)} (\operatorname{E}[Y \mid X = x] - \operatorname{E}[Y]) \cdot \operatorname{E}[Y - \operatorname{E}[Y \mid X = x] \mid X = x] \cdot \operatorname{P}(X = x)\\
&= 0
\end{aligned}$$

Continuous case.

$$\begin{aligned}\mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])] &= \int_{x\in\mathcal{R}(X)} \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Y\mid X]-\mathrm{E}[Y]) \mid X=x]\cdot \mathrm{p}(X=x)\,dx \\ &= \int_{x\in\mathcal{R}(X)} (\mathrm{E}[Y\mid X=x]-\mathrm{E}[Y])\cdot \mathrm{E}[Y-\mathrm{E}[Y\mid X=x]\mid X=x]\cdot \mathrm{p}(X=x)\,dx \\ &= 0\end{aligned}$$

Therefore:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}\left[(Y-\mathrm{E}[Y\mid X])^2\right] + \mathrm{E}\left[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])^2\right] \\ &= \mathrm{E}[\mathrm{Var}(Y\mid X)] + \mathrm{Var}(\mathrm{E}[Y\mid X])\end{aligned}$$

□

Definition 3.15 (Precision)**.** The **precision** of a random variable X, often denoted $\tau(X)$, τ_X, or shorthanded as τ, is the inverse of that random variable's variance; that is:

$$\tau(X) \stackrel{\text{def}}{=} (\mathrm{Var}(X))^{-1}$$

Definition 3.16 (Standard deviation)**.** The standard deviation of a random variable X is the square-root of the variance of X:

$$\mathrm{SD}(X) \stackrel{\text{def}}{=} \sqrt{\mathrm{Var}(X)}$$

Definition 3.17 (Covariance)**.** For any two one-dimensional random variables, X, Y:

$$\mathrm{Cov}(X,Y) \stackrel{\text{def}}{=} \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$$

Theorem 3.11 (Alternative formula for covariance)**.**

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

Theorem 3.12 (Law of total covariance)**.** *For random variables X, Y, and Z:*

$$\mathrm{Cov}(Y,Z) = \mathrm{E}[\mathrm{Cov}(Y,Z\mid X)] + \mathrm{Cov}(\mathrm{E}[Y\mid X], \mathrm{E}[Z\mid X])$$

where $\mathrm{Cov}(Y,Z\mid X) \stackrel{def}{=} \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(Z-\mathrm{E}[Z\mid X])\mid X]$.
Alternate names include: the ***covariance decomposition formula*** *and the* ***conditional covariance formula****.*

Proof

Proof. Write:

$$\begin{aligned}Y-\mathrm{E}[Y] &= (Y-\mathrm{E}[Y\mid X]) + (\mathrm{E}[Y\mid X]-\mathrm{E}[Y]) \\ Z-\mathrm{E}[Z] &= (Z-\mathrm{E}[Z\mid X]) + (\mathrm{E}[Z\mid X]-\mathrm{E}[Z])\end{aligned}$$

Then:

$$\begin{aligned}\operatorname{Cov}(Y,Z) &= \mathrm{E}[(Y-\mathrm{E}[Y])(Z-\mathrm{E}[Z])] \\ &= \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(Z-\mathrm{E}[Z\mid X])] \\ &\quad + \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z])] \\ &\quad + \mathrm{E}[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])(Z-\mathrm{E}[Z\mid X])] \\ &\quad + \mathrm{E}[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z])]\end{aligned}$$

For the two mixed terms:
Discrete case.

$$\begin{aligned}\mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z])] &= \sum_{x\in\mathcal{R}(X)} \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z]) \mid X=x]\cdot \mathrm{P}(X=x) \\ &= \sum_{x\in\mathcal{R}(X)} (\mathrm{E}[Z\mid X=x]-\mathrm{E}[Z])\cdot \mathrm{E}[Y-\mathrm{E}[Y\mid X=x]\mid X=x]\cdot \mathrm{P}(X=x) \\ &= 0\end{aligned}$$

and similarly:

$$\mathrm{E}[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])(Z-\mathrm{E}[Z\mid X])] = 0.$$

Continuous case.

$$\begin{aligned}\mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z])] &= \int_{x\in\mathcal{R}(X)} \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z]) \mid X=x]\cdot \mathrm{p}(X=x)\,dx \\ &= \int_{x\in\mathcal{R}(X)} (\mathrm{E}[Z\mid X=x]-\mathrm{E}[Z])\cdot \mathrm{E}[Y-\mathrm{E}[Y\mid X=x]\mid X=x]\cdot \mathrm{p}(X=x)\,dx \\ &= 0\end{aligned}$$

and similarly:

$$\mathrm{E}[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])(Z-\mathrm{E}[Z\mid X])] = 0.$$

Hence:

$$\begin{aligned}\operatorname{Cov}(Y,Z) &= \mathrm{E}[(Y-\mathrm{E}[Y\mid X])(Z-\mathrm{E}[Z\mid X])] + \mathrm{E}[(\mathrm{E}[Y\mid X]-\mathrm{E}[Y])(\mathrm{E}[Z\mid X]-\mathrm{E}[Z])] \\ &= \mathrm{E}[\operatorname{Cov}(Y,Z\mid X)] + \operatorname{Cov}(\mathrm{E}[Y\mid X],\mathrm{E}[Z\mid X])\end{aligned}$$

□

Lemma 3.1 (The covariance of a variable with itself is its variance)**.** *For any random variable X:*

$$\operatorname{Cov}(X,X) = \operatorname{Var}(X)$$

Proof

Proof.

$$\begin{aligned}\operatorname{Cov}(X,X) &= \mathrm{E}[XX] - \mathrm{E}[X]\,\mathrm{E}[X] \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\ &= \operatorname{Var}(X)\end{aligned}$$

□

Definition 3.18 (Variance/covariance of a $p \times 1$ random vector)**.** For a $p \times 1$ dimensional random vector $\tilde{X}$,

$$\begin{aligned}\operatorname{Var}\left(\tilde{X}\right) &\stackrel{\text{def}}{=} \operatorname{Cov}\left(\tilde{X}\right) \\ &\stackrel{\text{def}}{=} \mathrm{E}\left[\left(\tilde{X} - \mathrm{E}\,\tilde{X}\right)\left(\tilde{X} - \mathrm{E}\,\tilde{X}\right)^{\top}\right]\end{aligned}$$

Theorem 3.13 (Elements of the variance-covariance matrix are pairwise covariances)**.** *For a $p \times 1$ random vector $\tilde{X} = (X_1, \ldots, X_p)^{\top}$, the (i,j)-th element of $\operatorname{Var}\left(\tilde{X}\right)$ is $\operatorname{Cov}(X_i, X_j)$:*

$$\operatorname{Var}\left(\tilde{X}\right) = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \operatorname{Cov}(X_p, X_2) & \cdots & \operatorname{Var}(X_p) \end{pmatrix}$$

Proof

Proof. Let $\mu_i = \mathrm{E}[X_i]$ for $i = 1, \ldots, p$, so $\mathrm{E}\,\tilde{X} = (\mu_1, \ldots, \mu_p)^{\top}$. By Definition 3.18:

$$\begin{aligned}\operatorname{Var}\left(\tilde{X}\right) &= \mathrm{E}\left[\left(\tilde{X} - \mathrm{E}\,\tilde{X}\right)\left(\tilde{X} - \mathrm{E}\,\tilde{X}\right)^{\top}\right] \\ &= \mathrm{E}\left[\begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & \cdots & X_p - \mu_p \end{pmatrix}\right] \\ &= \mathrm{E}\left[\begin{pmatrix} (X_1 - \mu_1)(X_1 - \mu_1) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & \cdots & (X_p - \mu_p)(X_p - \mu_p) \end{pmatrix}\right] \\ &= \begin{pmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathrm{E}\left[(X_1 - \mu_1)(X_p - \mu_p)\right] \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left[(X_p - \mu_p)(X_1 - \mu_1)\right] & \cdots & \mathrm{E}\left[(X_p - \mu_p)(X_p - \mu_p)\right] \end{pmatrix} \\ &= \begin{pmatrix} \operatorname{Cov}(X_1, X_1) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \cdots & \operatorname{Cov}(X_p, X_p) \end{pmatrix} \\ &= \begin{pmatrix} \operatorname{Var}(X_1) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \cdots & \operatorname{Var}(X_p) \end{pmatrix}\end{aligned}$$

where:

- the step from the third to fourth line uses Definition 3.11,
- the step from the fourth to fifth line uses Definition 3.17, and
- the last step uses Lemma 3.1.

□

Theorem 3.14 (Alternate expression for variance of a random vector)**.**

$$\operatorname{Var}\left(\tilde{X}\right) = \mathrm{E}\left[\tilde{X}\tilde{X}^{\top}\right] - \left(\mathrm{E}\,\tilde{X}\right)\left(\mathrm{E}\,\tilde{X}\right)^{\top}$$

Proof

Proof.

$$\begin{aligned}
\operatorname{Var}\big(\tilde{X}\big) &= \operatorname{E}\Big[\big(\tilde{X} - \operatorname{E}\tilde{X}\big)\big(\tilde{X} - \operatorname{E}\tilde{X}\big)^\top\Big] \\
&= \operatorname{E}\Big[\tilde{X}\tilde{X}^\top - \tilde{X}\big(\operatorname{E}\tilde{X}\big)^\top - \big(\operatorname{E}\tilde{X}\big)\tilde{X}^\top + \big(\operatorname{E}\tilde{X}\big)\big(\operatorname{E}\tilde{X}\big)^\top\Big] \\
&= \operatorname{E}\big[\tilde{X}\tilde{X}^\top\big] - \big(\operatorname{E}\tilde{X}\big)\big(\operatorname{E}\tilde{X}\big)^\top - \big(\operatorname{E}\tilde{X}\big)\big(\operatorname{E}\tilde{X}\big)^\top + \big(\operatorname{E}\tilde{X}\big)\big(\operatorname{E}\tilde{X}\big)^\top \\
&= \operatorname{E}\big[\tilde{X}\tilde{X}^\top\big] - \big(\operatorname{E}\tilde{X}\big)\big(\operatorname{E}\tilde{X}\big)^\top
\end{aligned}$$

□

Theorem 3.15 (Variance of a linear combination)**.** *For any vector of random variables* $\tilde{X} = (X_1, \ldots, X_n)$ *and corresponding vector of constants* $\tilde{a} = (a_1, \ldots, a_n)$*, the variance of their linear combination is:*

$$\begin{aligned}
\operatorname{Var}\big(\tilde{a} \cdot \tilde{X}\big) &= \operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) \\
&= \tilde{a}^\top \operatorname{Var}\big(\tilde{X}\big)\tilde{a} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j)
\end{aligned}$$

Proof

Proof. Left to the reader… □

Corollary 3.2 (Variance of a sum of two random variables)**.** *For any two random variables* X *and* Y *and scalars* a *and* b*:*

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2(a \cdot b) \operatorname{Cov}(X, Y)$$

Proof

Proof. Apply Theorem 3.15 with $n = 2$, $X_1 = X$, and $X_2 = Y$.
Or, see https://statproofbook.github.io/P/var-lincomb.html □

Definition 3.19 (Homoskedastic and heteroskedastic)**.** A random variable Y is **homoskedastic** (with respect to covariates X) if the variance of Y does not vary with X:

$$\operatorname{Var}(Y|X = x) = \sigma^2, \forall x$$

Otherwise it is **heteroskedastic**.

Definition 3.20 (Statistical independence)**.** A set of random variables $X_1, \dots, X_n$ are **statistically independent** if their joint probability is equal to the product of their marginal probabilities:

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \Pr(X_i = x_i)$$

Tip

The symbol for independence, $\perp\!\!\!\perp$, is essentially just $\prod$ upside-down. So the symbol can remind you of its definition (Definition 3.20).

Definition 3.21 (Conditional independence)**.** A set of random variables $Y_1, \dots, Y_n$ are **conditionally statistically independent** given a set of covariates $X_1, \dots, X_n$ if the joint probability of the Y_is given the X_is is equal to the product of their marginal probabilities:

$$\Pr(Y_1 = y_1, \dots, Y_n = y_n | X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \Pr(Y_i = y_i | X_i = x_i)$$

Definition 3.22 (Identically distributed)**.** A set of random variables $X_1, \dots, X_n$ are **identically distributed** if they have the same range $\mathcal{R}(X)$ and if their marginal distributions $\mathrm{P}(X_1 = x_1), ..., \mathrm{P}(X_n = x_n)$ are all equal to some shared distribution $\mathrm{P}(X = x)$:

$$\forall i \in \{1 : n\}, \forall x \in \mathcal{R}(X) : \mathrm{P}(X_i = x) = \mathrm{P}(X = x)$$

Definition 3.23 (Conditionally identically distributed)**.** A set of random variables $Y_1, \dots, Y_n$ are **conditionally identically distributed** given a set of covariates $X_1, \dots, X_n$ if $Y_1, \dots, Y_n$ have the same range $\mathcal{R}(X)$ and if the distributions $\mathrm{P}(Y_i = y_i | X_i = x_i)$ are all equal to the same distribution $\mathrm{P}(Y = y | X = x)$:

$$\mathrm{P}(Y_i = y | X_i = x) = \mathrm{P}(Y = y | X = x)$$

Definition 3.24 (Independent and identically distributed)**.** A set of random variables $X_1, \dots, X_n$ are **independent and identically distributed** (shorthand: "X_i iid") if they are statistically independent and identically distributed.

Definition 3.25 (Conditionally independent and identically distributed)**.** A set of random variables $Y_1, \dots, Y_n$ are **conditionally independent and identically distributed** (shorthand: "$Y_i | X_i$ ciid" or just "$Y_i | X_i$ iid") given a set of covariates $X_1, \dots, X_n$ if $Y_1, \dots, Y_n$ are conditionally independent given $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are conditionally identically distributed given $X_1, \dots, X_n$.

3.7 The Central Limit Theorem

The sum of many independent or nearly-independent random variables with small variances (relative to the number of RVs being summed) produces bell-shaped distributions.

For example, consider the sum of five dice (Figure 8).

```
library(dplyr)
dist =
  expand.grid(1:6, 1:6, 1:6, 1:6, 1:6) |>
  rowwise() |>
```

```
  mutate(total = sum(c_across(everything()))) |>
  ungroup() |>
  count(total) |>
  mutate(`p(X=x)` = n/sum(n))

library(ggplot2)

dist |>
  ggplot() +
  aes(x = total, y = `p(X=x)`) +
  geom_col() +
  xlab("sum of dice (x)") +
  ylab("Probability of outcome, Pr(X=x)") +
  expand_limits(y = 0)
```

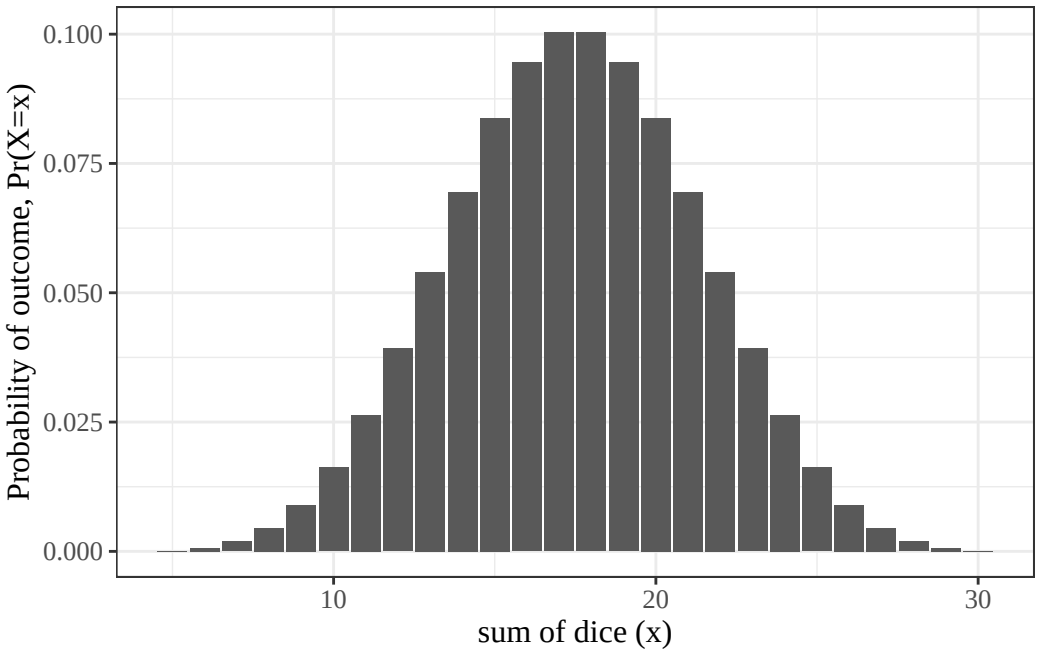


Figure 8: Distribution of the sum of five dice

In comparison, the outcome of just one die is not bell-shaped (Figure 9).

```
library(dplyr)
dist =
  expand.grid(1:6) |>
  rowwise() |>
  mutate(total = sum(c_across(everything()))) |>
  ungroup() |>
  count(total) |>
  mutate(`p(X=x)` = n/sum(n))

library(ggplot2)

dist |>
  ggplot() +
```

```
aes(x = total, y = `p(X=x)`) +
geom_col() +
xlab("sum of dice (x)") +
ylab("Probability of outcome, Pr(X=x)") +
expand_limits(y = 0)
```

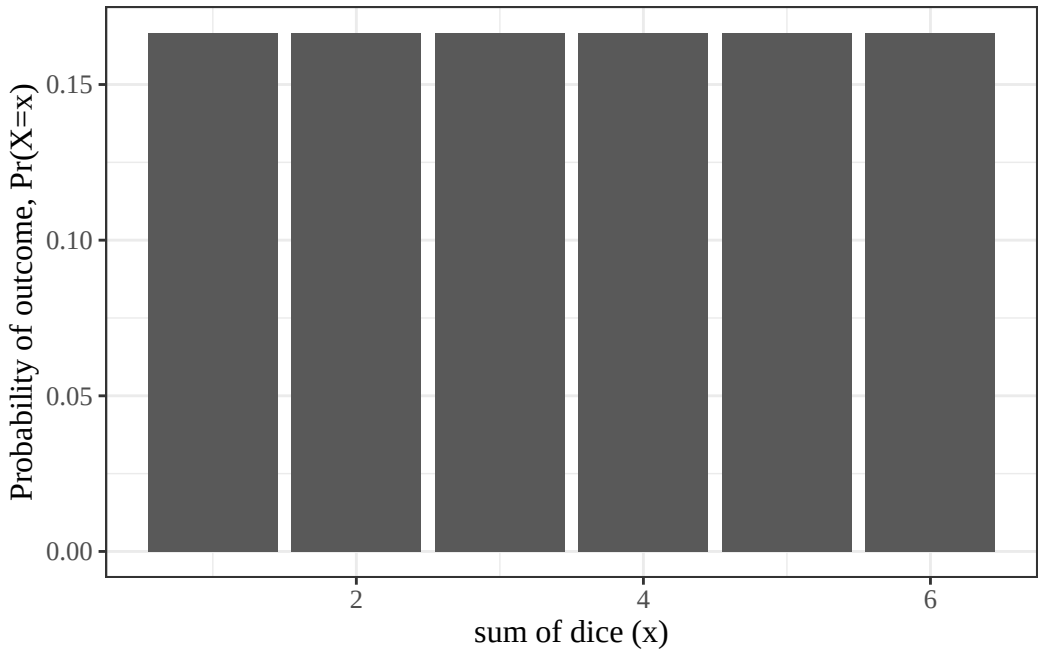


Figure 9: Distribution of the outcome of one die

What distribution does a single die have?

Answer: discrete uniform on 1:6.

4 Additional resources

- Miller (2017)

References

Billingsley, Patrick. 1995. *Probability and Measure*. 3rd ed. Wiley Series in Probability and Mathematical Statistics. Wiley.

Casella, George, and Roger Berger. 2002. *Statistical Inference*. 2nd ed. Cengage Learning. https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Gut, Allan. 2013. *Probability: A Graduate Course*. 2nd ed. Springer Texts in Statistics. Springer. https://doi.org/10.1007/978-1-4614-4708-5.

Kalbfleisch, John D, and Ross L Prentice. 2011. *The Statistical Analysis of Failure Time Data*.

John Wiley & Sons.

Klein, John P, and Melvin L Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. Vol. 1230. Springer. https://link.springer.com/book/10.1007/b97377.

Kleinbaum, David G, and Mitchel Klein. 2012. *Survival Analysis: A Self-Learning Text*. 3rd ed. Springer. https://link.springer.com/book/10.1007/978-1-4419-6646-9.

Miller, Steven J. 2017. *The Probability Lifesaver : All the Tools You Need to Understand Chance*. A Princeton Lifesaver Study Guide. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691149547/the-probability-lifesaver.

Rothman, Kenneth J., Timothy L. Lash, Tyler J. VanderWeele, and Sebastien Haneuse. 2021. *Modern Epidemiology*. Fourth edition. Wolters Kluwer.

Rudin, Walter. 1976. *Principles of Mathematical Analysis*. 3rd ed. International Series in Pure and Applied Mathematics. McGraw-Hill.

Soch, Joram. 2020. *Proof: Expected Value of a Non-Negative Random Variable*. The Book of Statistical Proofs. https://doi.org/10.5281/zenodo.4305949.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.

Wikipedia contributors. 2026. *Law of the Unconscious Statistician — Wikipedia, the Free Encyclopedia*. https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician.