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
library(haven) # import Stata files
```

```r
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. □

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:*

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

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

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:

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

□

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:

$$\begin{aligned} \Pr(A) &= \sum_{i=1}^{\infty} \Pr(A \cap B_i) \\ &= \sum_{i=1}^{\infty} \Pr(A \mid B_i) \cdot \Pr(B_i) \end{aligned}$$

□

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

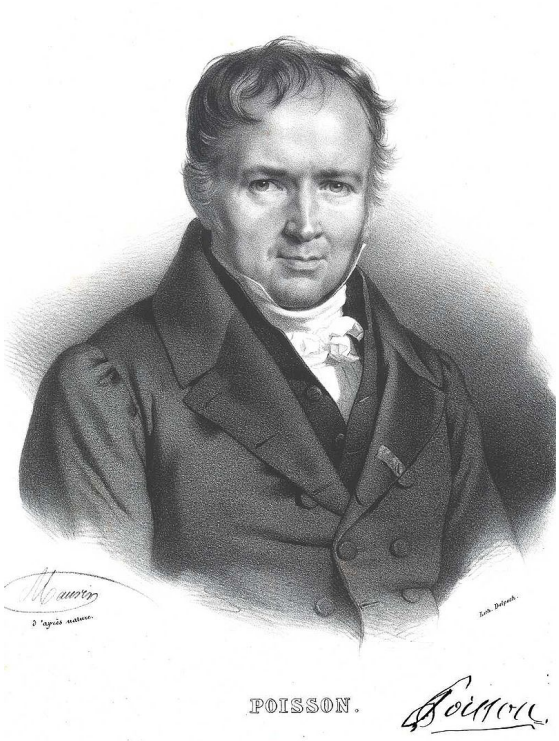
Distribution	Uses
χ^2	Regression comparisons (asymptotic), contingency table independence tests, goodness-of-fit tests
F	Gaussian model comparisons (exact)
Z (standard normal)	Proportions, means, regression coefficients (asymptotic)
T	Means, regression coefficients in Gaussian outcome models (exact)

2.1 The Bernoulli distribution

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

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

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).

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

(see Figure 2)

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

Solution

Solution 2.2.

$$\mathcal{R}(Y) = \{0, 1, 2, ...\} = \mathbb{N}$$

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

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

(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)
```

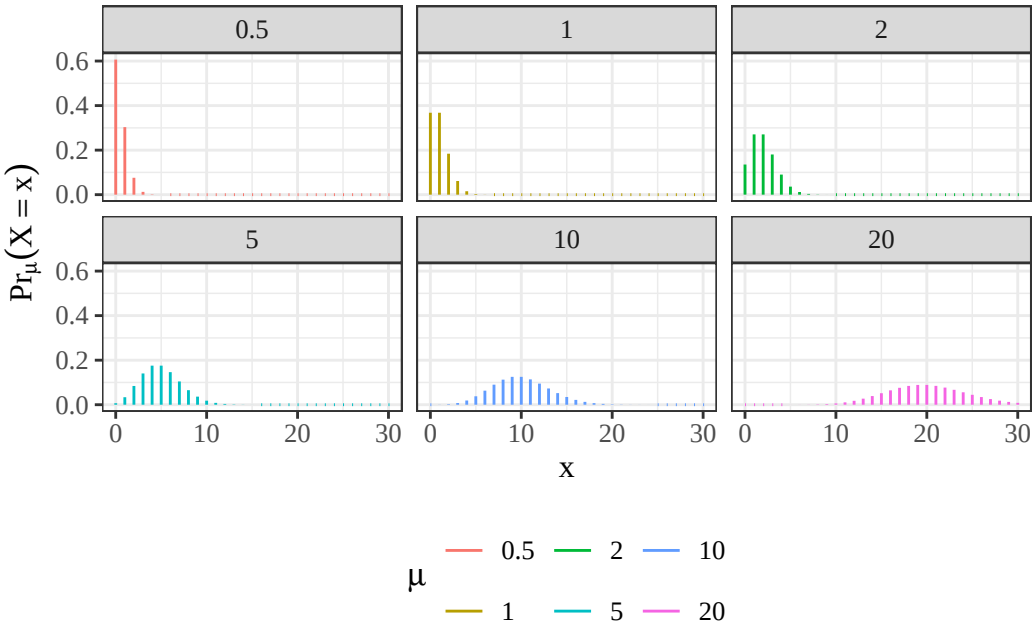


Figure 2: Poisson PMFs, by mean parameter μ

```
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)
```

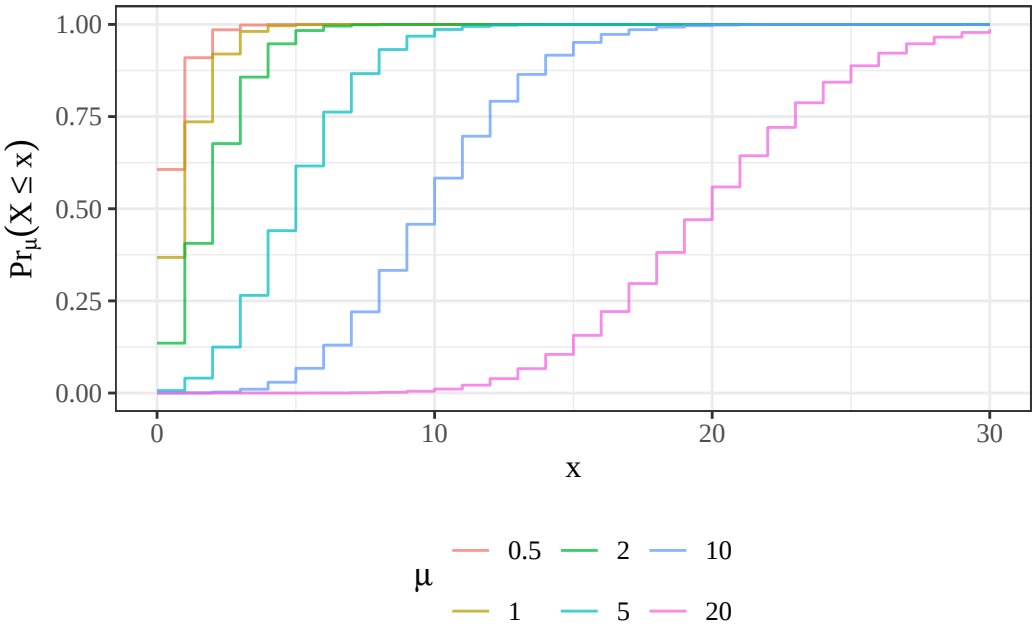


Figure 3: Poisson CDFs

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.

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_{x=0}^{\infty} x \cdot P(X = x) \\
&= 0 \cdot P(X = 0) + \sum_{x=1}^{\infty} x \cdot P(X = x) \\
&= 0 + \sum_{x=1}^{\infty} x \cdot P(X = x) \\
&= \sum_{x=1}^{\infty} x \cdot P(X = x) \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x \cdot (x-1)!} && \text{[definition of factorial (”!”) function]} \\
&= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&= \sum_{x=1}^{\infty} \frac{(\lambda \cdot \lambda^{x-1}) e^{-\lambda}}{(x-1)!} \\
&= \lambda \cdot \sum_{x=1}^{\infty} \frac{(\lambda^{x-1}) e^{-\lambda}}{(x-1)!} \\
&= \lambda \cdot \sum_{y=0}^{\infty} \frac{(\lambda^{y}) e^{-\lambda}}{(y)!} && \text{[substituting } y \stackrel{\text{def}}{=} x - 1\text{]} \\
&= \lambda \cdot 1 && \text{[because PDFs sum to 1]} \\
&= \lambda
\end{aligned}
$$

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. That is:

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

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

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:*

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

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.

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

□

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:*

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

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. See https://web.stanford.edu/class/archive/cs/cs109/cs109.1206/lectureNotes/LN12_independent_rvs.pdf, Example 3. □

2.3 The Negative-Binomial distribution

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

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

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* $Y \sim \text{NegBin}(\mu, \rho)$, *then:*

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

2.4 Weibull Distribution

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

When $\alpha = 1$ this is the exponential. When $\alpha > 1$ the hazard is increasing and when $\alpha < 1$ the hazard is decreasing. This 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 $f(x)$, $f_X(x)$, $\text{p}(x)$, $\text{p}_X(x)$, or $\text{p}(X = x)$, 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.

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

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

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

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

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

Theorem 3.1 (Density function is derivative of CDF). *The density function* $f(t)$ *or* $\text{p}(T = t)$ *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 $T \geq t$. That is:

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

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 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] 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)

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:

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

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:

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

(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.

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

□

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

$$\mu(T|\tilde{X} = \tilde{x}) = \int_{t=0}^{\infty} \mathrm{S}(t) dt$$

Proof

Proof. We prove the continuous case, in which T has a density f. The result follows from applying Tonelli's theorem (hypothesis (a) of Fubini–Tonelli[a]) to the function $g(t, u) = \mathrm{f}(u) \cdot \mathbb{1}(0 \leq t \leq u)$ on the product space $[0, \infty) \times [0, \infty)$: g is nonnegative everywhere and vanishes outside the (unbounded) triangular region $D = \{(t, u) : 0 \leq t \leq u < \infty\}$, so the

iterated integrals over D are exchangeable.
Since $\mathrm{f}(u) \geq 0$, hypothesis (a) of Fubini–Tonelli[b] (the nonnegative case, **Tonelli's theorem**) applies, and we may exchange the order of integration over D:

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

□

[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))**.** *For any function g of a* discrete *random variable X:*

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

Proof

Proof. Let $Y = g(X)$. By Definition 3.5 applied to Y:

$$\begin{aligned}
\mathrm{E}[g(X)] &= \mathrm{E}[Y] \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(Y = y) \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathrm{P}(g(X) = y) \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{\substack{x \in \mathcal{R}(X) \\ g(x) = y}} \mathrm{P}(X = x) \\
&= \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathrm{P}(X = x)
\end{aligned}$$

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

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.

For a *continuous* random variable X with density $\mathrm{p}(X = x)$, the analogous formula is:

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

Exm

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

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

Definition 3.6 (Conditional expectation)**. Discrete case.** 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:

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

The **conditional expectation** of Y given $X = x$ is:

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

Continuous case. Let X and Y be jointly distributed continuous random variables 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:

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

The **conditional expectation** of Y given $X = x$ is:

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

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; that is, $\mathrm{E}[Y \mid X] = g(X)$ where $g(x) \overset{\text{def}}{=} \mathrm{E}[Y \mid X = x]$.

3.4 Fubini–Tonelli for expectations

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

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 hypothesis 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#thm-fubini-tonelli
[6] math-prereqs.qmd#thm-fubini-tonelli

Exm

Example 3.3 (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 3.4 (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:

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

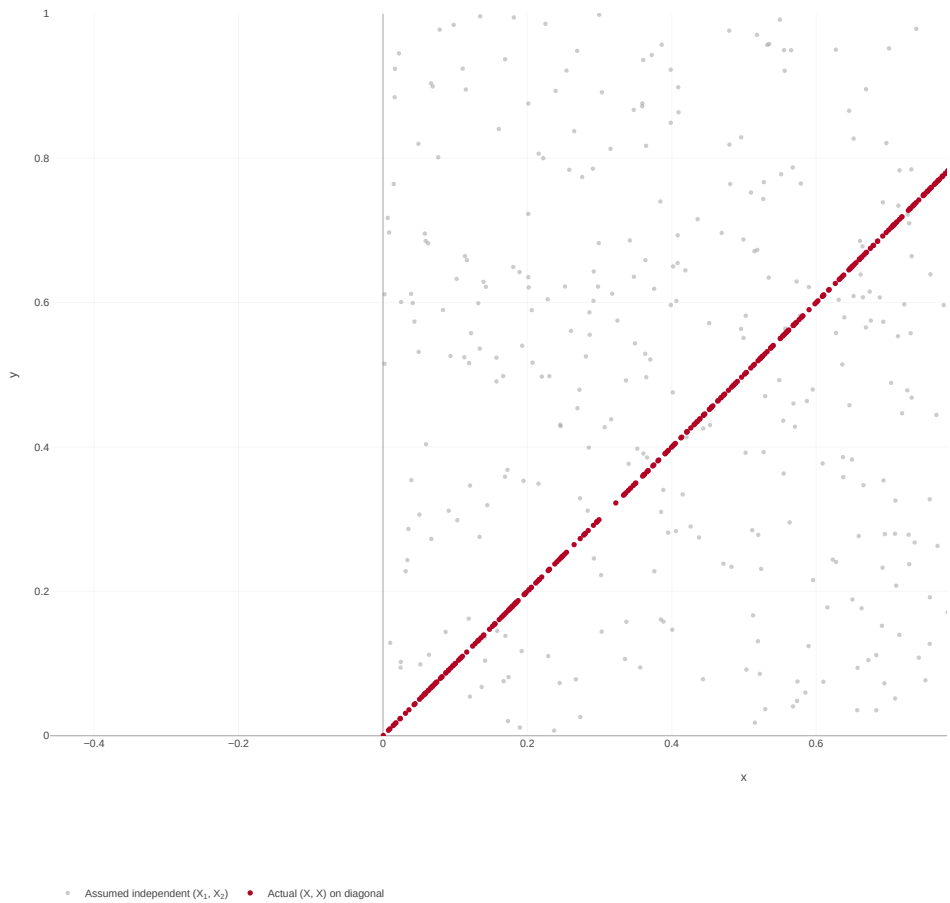
Erroneously applying the product-measure formula:
Note that Fubini–Tonelli's own conditions still hold here ($h(x,y) = xy$ is nonnegative and integrable), so the error is not a failure of Fubini–Tonelli[a]. Rather, the error is using the *wrong measure*: the joint distribution of (X,X) is concentrated 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 calculation below 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 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 3.5 (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))
  ))
```

Exm

Example 3.6 (Both-discrete case: 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 $\mathrm{E}[X+Y]$ using Corollary 3.1 with $\mu_X = \mu_Y =$ counting measure and $h(x,y) = x+y$. By Corollary 3.1 (both-discrete case):

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

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 $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 1.2$.
Note that X and Y are **not** independent here: $\mathrm{P}(X=0, Y=0) = 0.2 \neq 0.15 = \mathrm{P}(X=0)\,\mathrm{P}(Y=0)$. Corollary 3.1 applies regardless, since it requires only the *actual* joint mass function, not independence.

```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"))
  )
```

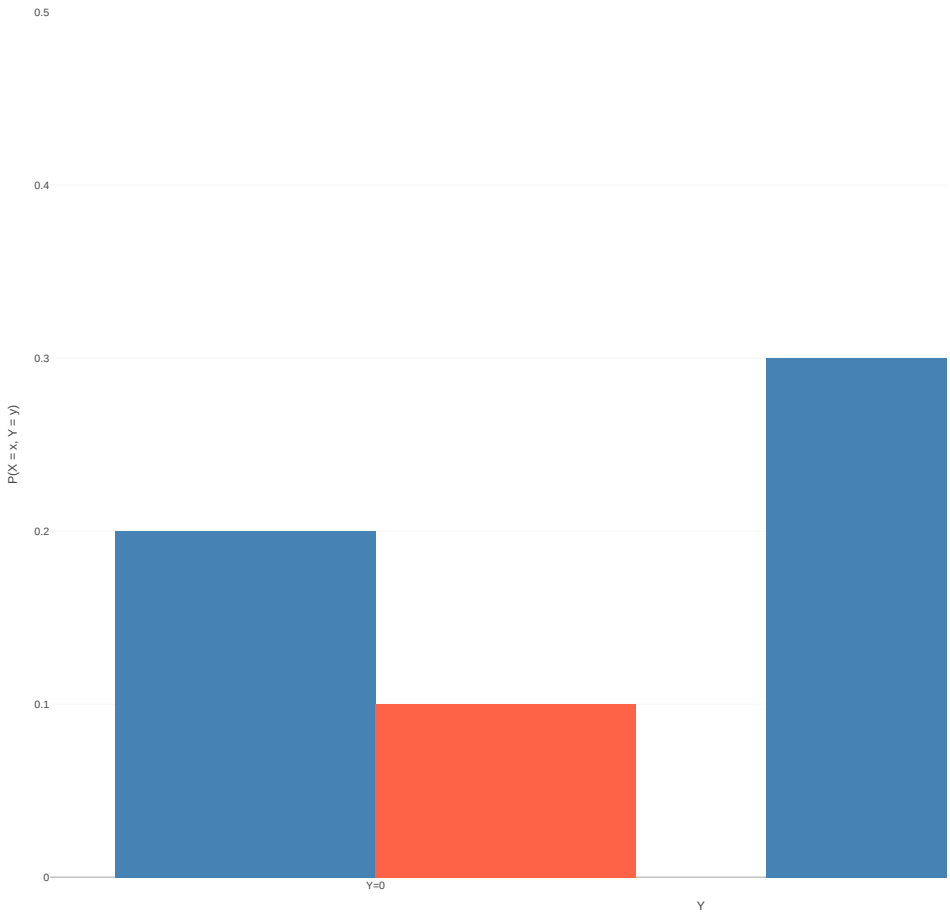


Figure 6: Joint probability mass function $\mathrm{P}(X = x, Y = y)$. Marginal totals: $\mathrm{P}(X = 0) = 0.5$, $\mathrm{P}(X = 1) = 0.5$, $\mathrm{P}(Y = 0) = 0.3$, $\mathrm{P}(Y = 1) = 0.7$.

Exm

Example 3.7 (Mixed case: one continuous variable, one discrete variable)**.** Let $Y \sim \text{Bernoulli}(0.6)$ and, given $Y = y$, let $X \mid Y = y \sim \text{Uniform}(0,\, y+1)$. Compute $\mathrm{E}[X]$ using Corollary 3.1 with $\mu_X =$ Lebesgue measure, $\mu_Y =$ counting measure, and $h(x, y) = x$. The joint density w.r.t. Lebesgue $\times$ counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\ \mathrm{P}(Y = y)$:

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

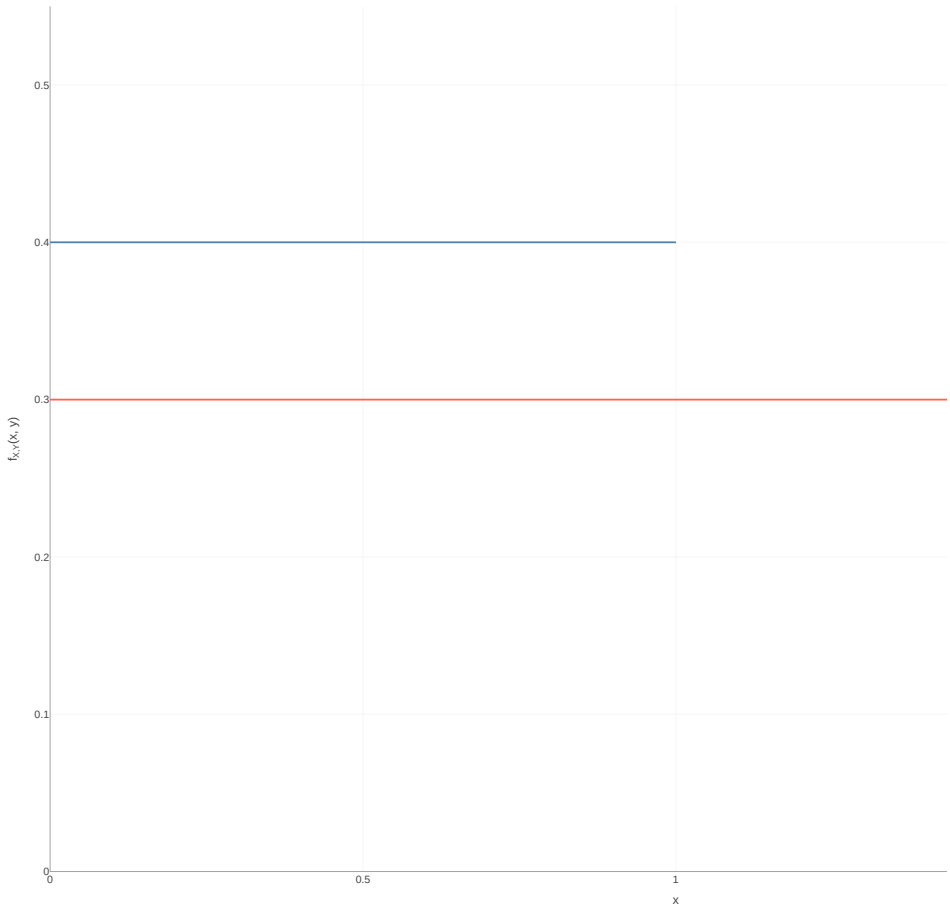
By Corollary 3.1 (mixed case):

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

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"))
  )
```



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 $\{X = x : x \in \mathcal{R}(X)\}$:

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

Continuous case. When X and Y are continuous, applying Definition 3.5 to $\mathrm{E}[\mathrm{E}[Y \mid X]]$ and then using Definition 3.6 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 hypothesis (b) of Fubini–Tonelli[a] (the absolute-integrability case, **Fubini's theorem**); this 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 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 $Z = z$:

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

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 3.8 (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 3.7 (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 3.8 (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:

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

Definition 3.9 (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 3.10 (Variance)**.** The variance of a random variable X is the expectation of the squared difference between X and $\mathrm{E}[X]$; that is:

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

Theorem 3.8 (Simplified expression for variance)**.**

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

Proof

Proof. By linearity of expectation, we have:

$$\begin{aligned}\operatorname{Var}(X) &\overset{\text{def}}{=} \mathrm{E}[(X-\mathrm{E}[X])^2] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\ &= \mathrm{E}[X^2] - \mathrm{E}[2X\,\mathrm{E}[X]] + \mathrm{E}\left[(\mathrm{E}[X])^2\right] \\ &= \mathrm{E}[X^2] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2\end{aligned}$$

□

Theorem 3.9 (Law of total variance)**.** *For random variables X and Y:*

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

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

Proof

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

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

Taking expectation:

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

For the cross-term:
Discrete case.

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

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:

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

□

Definition 3.11 (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:

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

Definition 3.12 (Standard deviation)**.** The standard deviation of a random variable X is the square-root of the variance of X:

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

Definition 3.13 (Covariance)**.** For any two one-dimensional random variables, X, Y:

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

Theorem 3.10 (Alternative formula for covariance)**.**

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

Theorem 3.11 (Law of total covariance)**.** *For random variables X, Y, and Z:*

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

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

Proof

Proof. Write:

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

Then:

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

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:

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

□

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

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

Proof

Proof.

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

□

Definition 3.14 (Variance/covariance of a $p\times 1$ random vector)**.** For a $p\times 1$ dimensional random vector $\tilde{X}$,

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

Theorem 3.12 (Elements of the variance-covariance matrix are pairwise covariances)**.** *For a $p \times 1$ random vector $\tilde{X} = (X_1, \dots, X_p)^\top$, the (i,j)-th element of $\mathrm{Var}\big(\tilde{X}\big)$ is $\mathrm{Cov}(X_i, X_j)$:*

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

Proof

Proof. Let $\mu_i = \mathrm{E}[X_i]$ for $i = 1, \dots, p$, so $\mathrm{E}\,\tilde{X} = (\mu_1, \dots, \mu_p)^\top$. By Definition 3.14:

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

where:

- the step from the third to fourth line uses Definition 3.7,
- the step from the fourth to fifth line uses Definition 3.13, and
- the last step uses Lemma 3.1.

□

Theorem 3.13 (Alternate expression for variance of a random vector)**.**

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

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

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

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

Definition 3.15 (homoskedastic, 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 3.16 (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 3.16).

Definition 3.17 (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 3.18 (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 3.19 (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 3.20 (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 3.21 (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 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)
```

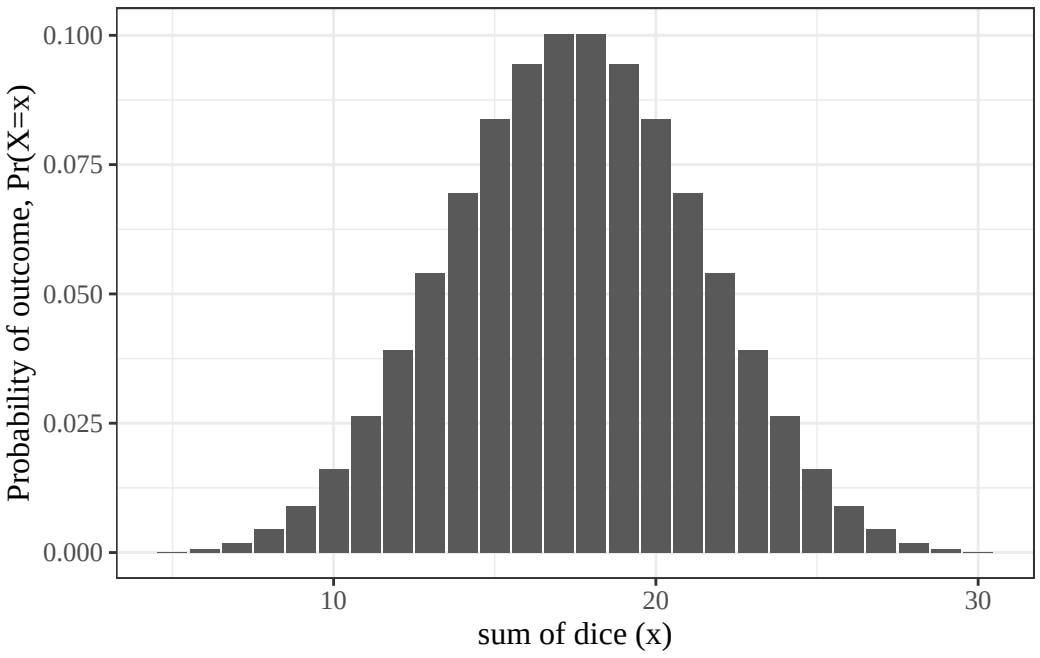


Figure 8: Distribution of the sum of five dice

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

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

```r
  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)
```

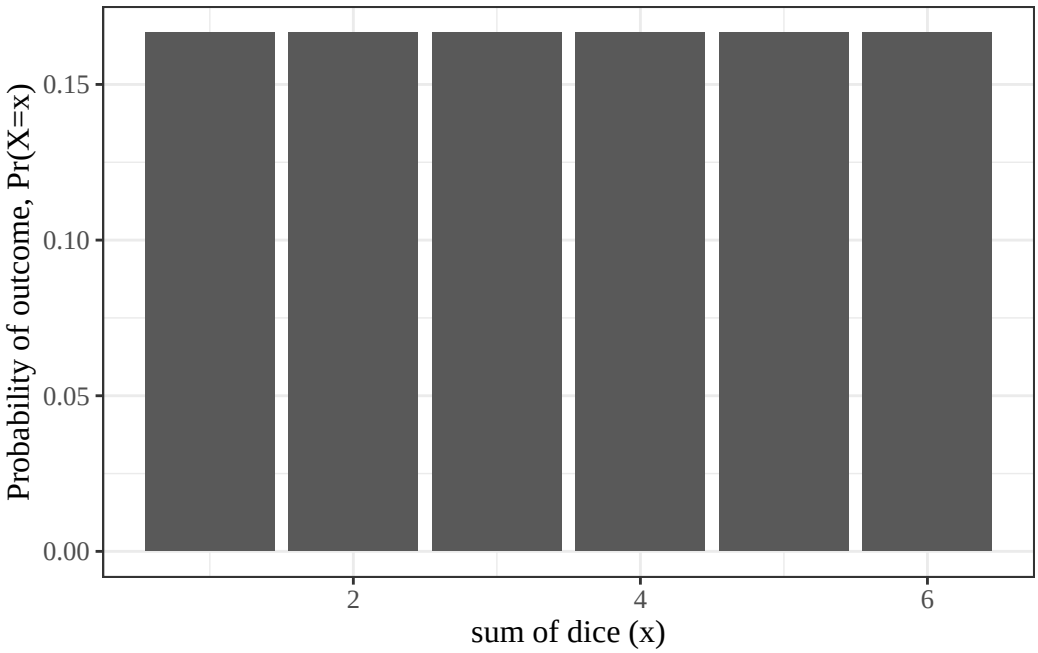


Figure 9: Distribution of the outcome of one die

What distribution does a single die have?

Answer: discrete uniform on 1:6.

4 Additional resources

- Miller (2017)

References

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

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.

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.