

Probability

Contents

Configuring R	1
1 Core properties of probabilities	2
1.1 Defining probabilities	2
1.2 Conditional probability	3
2 Random variables	5
2.1 Binary variables	5
2.2 Count variables	5
3 Key probability distributions	7
3.1 The Bernoulli distribution	8
3.2 The Poisson distribution	9
3.3 The Negative-Binomial distribution	14
3.4 Weibull Distribution	14
4 Characteristics of probability distributions	14
4.1 Probability density function	14
4.2 Hazard function	15
4.3 Expectation	16
4.4 Variance and related characteristics	20
4.5 The Central Limit Theorem	25
5 Additional resources	27
References	27

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(tidy) # 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
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

```

Here are some R settings I use in this document:

```

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 -algebra¹ 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, \dots (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*):

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

¹<https://en.wikipedia.org/wiki/%CE%A3-algebra>

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

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

for every finite collection of mutually disjoint events. Countable additivity implies finite additivity (set $A_{n+1} = A_{n+2} = \dots = \emptyset$ 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 σ -additive²). Requiring countable additivity enables results such as the continuity of probability (if $A_1 \supseteq A_2 \supseteq \dots$ with $\bigcap_i A_i = \emptyset$, then $\Pr(A_i) \rightarrow 0$) and underpins the Theorem 1.4 for countable partitions.

Theorem 1.1. *If A and B are statistical events and $A \subseteq B$, then $\Pr(A \cap B) = \Pr(A)$.*

Proof. Left to the reader for now. □

Theorem 1.2.

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

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

Corollary 1.1.

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

Proof. By Theorem 1.2 and algebra. □

Corollary 1.2. *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. Using Corollary 1.1:

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

□

²https://en.wikipedia.org/wiki/Sigma-additive_set_function#An_additive_function_which_is_not_%CF%83-additive

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 $\Pr(A | B)$, is:

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

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

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

Proof. Rearranging Definition 1.2:

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

□

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 ($\Pr(L | E) = 0.60$).

Then the probability that a randomly selected adult both exercises regularly and has low blood pressure is:

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

Theorem 1.4 (Law of total probability). If B_1, B_2, \dots 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 :

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

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

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

□

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

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

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

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

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

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:

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

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

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

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 Random variables

2.1 Binary variables

Definition 2.1 (binary variable). A **binary variable** is a random variable which has only two possible values in its range.

Exercise 2.1 (Examples of binary variables). What are some examples of binary variables in the health sciences?

Solution. Examples of binary outcomes include:

- exposure (exposed vs unexposed)
- disease (diseased vs healthy)
- recovery (recovered vs unrecovered)

- relapse (relapse vs remission)
- return to hospital (returned vs not)
- vital status (dead vs alive)

2.2 Count variables

Definition 2.2 (Count variable). A **count variable** is a random variable whose possible values are some subset of the non-negative integers; that is, a random variable X such that:

$$\mathcal{R}(X) \in \mathbb{N}$$

Exercise 2.2. What are some examples of count variables?

Solution.

- Number of fish in a pond
- Number of cyclones per season
- Seconds of tooth-brushing per session (if rounded)³
- Infections per person-year
- Visits to ER per person-month
- Car accidents per 1000 miles driven

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 2.3. What are some examples of exposure magnitudes?

Solution.

Table 1: 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:

³<https://pubmed.ncbi.nlm.nih.gov/35587489/>

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

$$\lambda \stackrel{\text{def}}{=} \frac{\mu}{t} \tag{1}$$

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

$$\therefore \mu = \lambda \cdot t \tag{2}$$

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

Equation 2 is analogous to the inverse-odds function for binary variables.

Theorem 2.2. *When the exposure magnitude is 0, there is no opportunity for events to occur:*

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

Proof.

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

□

Probability distributions for count outcomes

- [Poisson distribution](#)
- [Negative binomial distribution](#)

3 Key probability distributions

Some distributions are typically used for outcome models (Table 2); other distributions are typically used for test statistics (Table 3).

Table 2: Distributions typically used for outcome models

Distribution	Uses
Bernoulli	Binary outcomes
Binomial	Sums of Bernoulli outcomes

Distribution	Uses
Poisson	unbounded count outcomes
Geometric	Counts of non-events before an event occurs
Negative binomial	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 3: 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)

3.1 The Bernoulli distribution

Definition 3.1 (Bernoulli distribution). The **Bernoulli distribution** family for a random variable X is defined as:

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

3.2 The Poisson distribution



(a) Siméon Denis Poisson



(b) Les Poissons^a

^a<https://youtu.be/UoJxBEQRLd0?t=12>

Figure 1: “Les Poissons”

Exercise 3.1. Define the Poisson distribution.

Solution 3.1.

Definition 3.2 (Poisson distribution).

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, y \in \mathbb{N} \quad (3)$$

(see Figure 2)

Exercise 3.2. What is the range of possible values for a Poisson distribution?

Solution 3.2.

$$\mathcal{X}(Y) = \{0, 1, 2, \dots\} = \mathbb{N}$$

Theorem 3.1 (CDF of Poisson distribution).

$$P(Y \leq y) = e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} \quad (4)$$

(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)

```

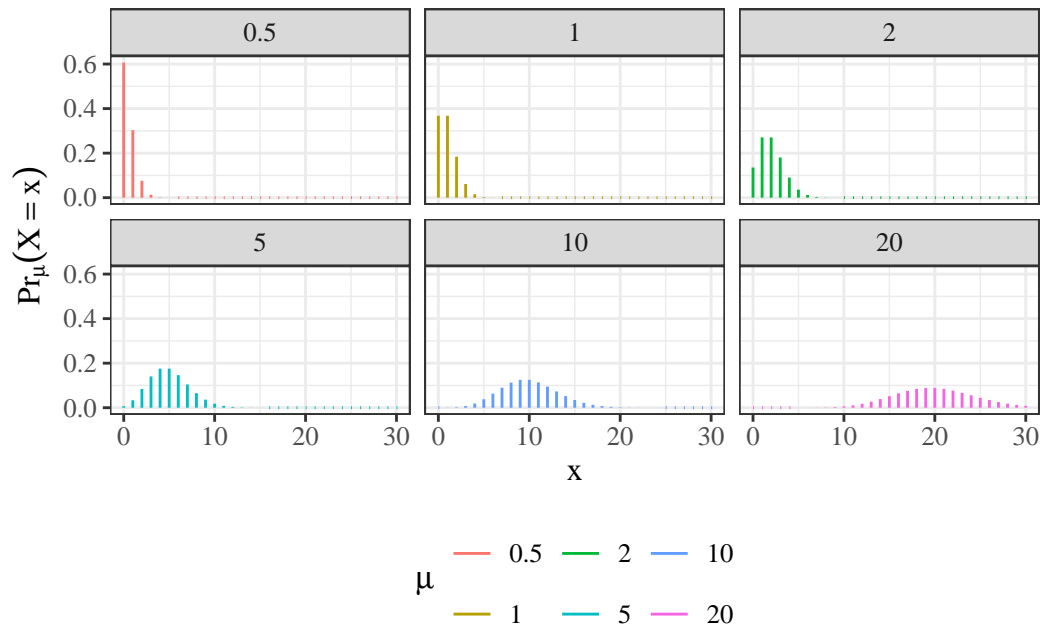


Figure 2: Poisson PMFs, by mean parameter μ

```
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)
```

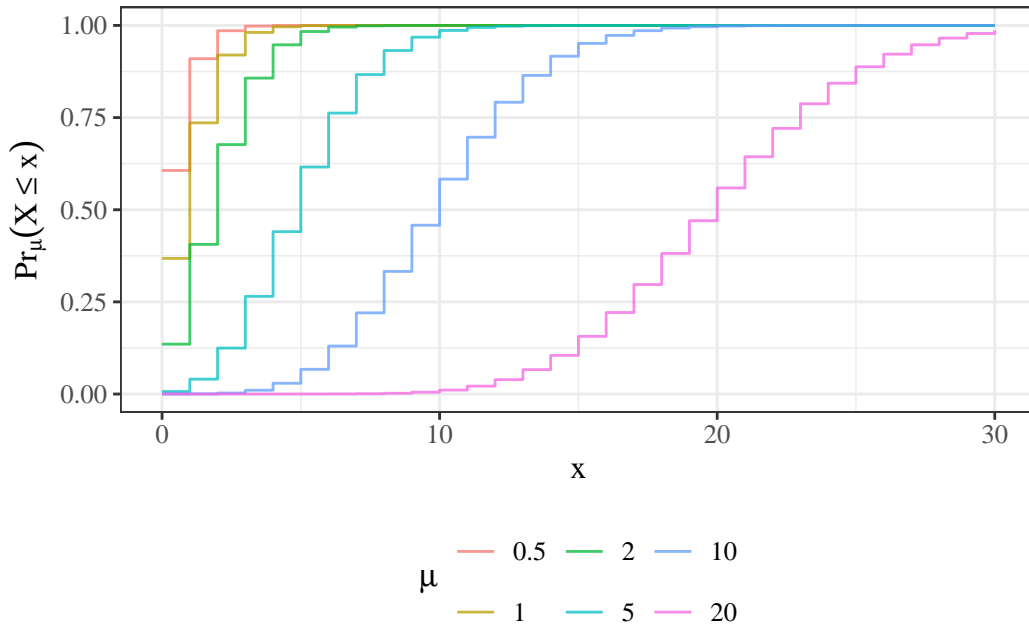


Figure 3: Poisson CDFs

Exercise 3.3 (Poisson distribution functions). Let $X \sim \text{Pois}(\mu = 3.75)$.

Compute:

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

Solution.

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

Theorem 3.2 (Properties of the Poisson distribution). *If $X \sim \text{Pois}(\mu)$, then:*

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

Exercise 3.4. Prove Theorem 3.2.

Solution.

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

See also <https://statproofbook.github.io/P/poiss-mean>.

For the variance, see <https://statproofbook.github.io/P/poiss-var>.

Accounting for exposure

If the exposures/observation durations, denoted $T = t$ or $N = n$, vary between observations, we model:

$$\mu = \lambda \cdot t$$

λ is interpreted as the “expected event rate per unit of exposure”; that is,

$$\lambda = \frac{\mathbb{E}[Y|T = t]}{t}$$

! 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 $\mathbb{E}[Y|T = 0]$** . In other words, this model assumes that if there is no exposure, there can’t be any events.

Theorem 3.3. *If $\mu = \lambda \cdot t$, then:*

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

Definition 3.3 (Offset). When the linear component of a model involves a term without an unknown coefficient, that term is called an **offset**.

Theorem 3.4. 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. See https://web.stanford.edu/class/archive/cs/cs109/cs109.1206/lectureNotes/LN12_independent_rvs.pdf, Example 3. □

3.3 The Negative-Binomial distribution

Definition 3.4 (Negative binomial distribution).

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

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

Theorem 3.5. If $Y \sim \text{NegBin}(\mu, \rho)$, then:

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

3.4 Weibull Distribution

$$p(t) = \alpha \lambda x^{\alpha-1} e^{-\lambda x^\alpha}$$

$$\lambda(t) = \alpha \lambda x^{\alpha-1}$$

$$S(t) = e^{-\lambda x^\alpha}$$

$$E(T) = \Gamma(1 + 1/\alpha) \cdot \lambda^{-1/\alpha}$$

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.

4 Characteristics of probability distributions

4.1 Probability density function

Definition 4.1 (probability density). If X is a continuous random variable, then the **probability density** of X at value x , denoted $f(x)$, $f_X(x)$, $p(x)$, $p_X(x)$, or $p(X = x)$, 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.

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

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

Theorem 4.1 (Density function is derivative of CDF). *The density function $f(t)$ or $p(T = t)$ for a random variable T at value t is equal to the derivative of the cumulative probability function $F(t) \stackrel{\text{def}}{=} P(T \leq t)$; that is:*

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

Theorem 4.2 (Density functions integrate to 1). *For any density function $f(x)$,*

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

4.2 Hazard function

Definition 4.2 (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 $h(t)$ ⁴ or $\lambda(t)$,⁵ is the conditional density⁶ of T at t , given $T \geq t$. That is:

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

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 .

Table 4: Probability distribution functions

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

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

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

⁴for example in Dobson and Barnett (2018), Vittinghoff et al. (2012), Klein and Moeschberger (2003), and Kleinbaum and Klein (2012)

⁵for example, in Rothman et al. (2021) and Kalbfleisch and Prentice (2011)

⁶[probability.qmd#def-pdf](#)

4.3 Expectation

Definition 4.3 (Expectation, expected value, population mean). The **expectation, expected value, or population mean** of a *continuous* random variable X , denoted $E[X]$, $\mu(X)$, or μ_X , is the weighted mean of X 's possible values, weighted by the probability density function of those values:

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

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

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

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

Theorem 4.3 (Expectation of the Bernoulli distribution). *The expectation of a Bernoulli random variable with parameter π is:*

$$E[X] = \pi$$

Proof.

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

□

Theorem 4.4 (Expectation of time-to-event variables). *If T is a non-negative random variable, then:*

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

Theorem 4.5 (Law of the Unconscious Statistician (LOTUS)). *For any function g of a discrete random variable X :*

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

Proof. Let $Y = g(X)$. By Definition 4.3 applied to Y :

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

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

LOTUS says that to compute $\mathbb{E}[g(X)]$, 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 $p(X = x)$, the analogous formula is:

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

Example 4.1 (Expected value of X^2 for a Bernoulli variable). Let $X \sim \text{Ber}(\pi)$. By LOTUS (Theorem 4.5):

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

Definition 4.4 (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 $\mathbb{P}(X = x) > 0$) is:

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

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

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

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

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

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

$$E[Y | X = x] \stackrel{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} y \cdot p(Y = y | X = x) dy$$

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

Theorem 4.6 (Law of iterated expectations). *For any two random variables X and Y :*

$$E[Y] = E[E[Y | X]]$$

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

Proof. Discrete case. When X and Y are discrete, applying Definition 4.3 to $E[E[Y | X]]$ and then the law of total probability (Theorem 1.4) applied to the countable partition $\{X = x : x \in \mathcal{R}(X)\}$:

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

Continuous case. When X and Y are continuous, applying Definition 4.3 to $E[E[Y | X]]$ and then using Definition 4.4 for $E[Y | X = x]$:

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

where the third equality exchanges the order of integration by Fubini's theorem, and the fourth equality uses $\int_x p(Y = y | X = x) \cdot p(X = x) dx = \int_x p(X = x, Y = y) dx = p(Y = y)$ (marginalization of the joint density). \square

Theorem 4.7 (Conditional law of iterated expectations). *For random variables X , Y , and Z :*

$$E[Y | Z] = E[E[Y | X, Z] | Z]$$

This is the tower rule applied conditionally on Z .

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 $Z = z$:

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

Continuous case. Conditioning on $Z = z$, and integrating over X under the conditional density $p(X = x | Z = z)$:

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

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

Example 4.2 (Marginal expectation from conditional expectations). Suppose X is a binary random variable indicating treatment assignment ($X = 1$ treated, $X = 0$ control), with $\mathbb{P}(X = 1) = 0.5$, and suppose the outcome Y has conditional expectations:

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

By the law of iterated expectations (Theorem 4.6):

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

Definition 4.5 (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** \mathbf{EA} is the $m \times n$ matrix whose (i, j) -th element is $\mathbb{E}[A_{ij}]$:

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

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

4.4 Variance and related characteristics

Definition 4.6 (Variance). The variance of a random variable X is the expectation of the squared difference between X and $E[X]$; that is:

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

Theorem 4.8 (Simplified expression for variance).

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

Proof. By linearity of expectation, we have:

$$\begin{aligned} \text{Var}(X) &\stackrel{\text{def}}{=} E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - E[2XE[X]] + E[(E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

□

Theorem 4.9 (Law of total variance). For random variables X and Y :

$$\text{Var}(Y) = E[\text{Var}(Y | X)] + \text{Var}(E[Y | X])$$

where $\text{Var}(Y | X) \stackrel{\text{def}}{=} E[(Y - E[Y | X])^2 | X]$.

Alternate names include: the **conditional variance formula**, **Eve's law**, and the **variance decomposition formula**.

Proof. Write $Y - E[Y] = (Y - E[Y | X]) + (E[Y | X] - E[Y])$. Then:

$$(Y - E[Y])^2 = (Y - E[Y | X])^2 + (E[Y | X] - E[Y])^2 + 2(Y - E[Y | X])(E[Y | X] - E[Y])$$

Taking expectation:

$$\begin{aligned} \text{Var}(Y) &= E[(Y - E[Y | X])^2] + E[(E[Y | X] - E[Y])^2] \\ &\quad + 2E[(Y - E[Y | X])(E[Y | X] - E[Y])] \end{aligned}$$

For the cross-term:

Discrete case.

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

Continuous case.

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

Therefore:

$$\begin{aligned} \text{Var}(Y) &= E[(Y - E[Y | X])^2] + E[(E[Y | X] - E[Y])^2] \\ &= E[\text{Var}(Y | X)] + \text{Var}(E[Y | X]) \end{aligned}$$

□

Definition 4.7 (Precision). The **precision** of a random variable X , often denoted $\tau(X)$, τ_X , or shorthand as τ , is the inverse of that random variable's variance; that is:

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

Definition 4.8 (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 4.9 (Covariance). For any two one-dimensional random variables, X, Y :

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

Theorem 4.10.

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

Theorem 4.11 (Law of total covariance). For random variables X, Y , and Z :

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

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

Alternate names include: the **covariance decomposition formula** and the **conditional covariance formula**.

Proof. Write:

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

Then:

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

For the two mixed terms:

Discrete case.

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

and similarly:

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

Continuous case.

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

and similarly:

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

Hence:

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

□

Lemma 4.1 (The covariance of a variable with itself is its variance). *For any random variable X :*

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

Proof.

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

□

Definition 4.10 (Variance/covariance of a $p \times 1$ random vector). For a $p \times 1$ dimensional random vector \tilde{X} ,

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

Theorem 4.12 (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 $\text{Var}(\tilde{X})$ is $\text{Cov}(X_i, X_j)$:

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

Proof. Let $\mu_i = \text{E}[X_i]$ for $i = 1, \dots, p$, so $\text{E}\tilde{X} = (\mu_1, \dots, \mu_p)^\top$. By Definition 4.10:

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

where:

- the step from the third to fourth line uses Definition 4.5,
- the step from the fourth to fifth line uses Definition 4.9, and
- the last step uses Lemma 4.1.

□

Theorem 4.13 (Alternate expression for variance of a random vector).

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

Proof.

$$\begin{aligned}
 \text{Var}(\tilde{X}) &= \mathbb{E}\left[(\tilde{X} - \mathbb{E}\tilde{X})(\tilde{X} - \mathbb{E}\tilde{X})^\top\right] \\
 &= \mathbb{E}\left[\tilde{X}\tilde{X}^\top - \tilde{X}(\mathbb{E}\tilde{X})^\top - (\mathbb{E}\tilde{X})\tilde{X}^\top + (\mathbb{E}\tilde{X})(\mathbb{E}\tilde{X})^\top\right] \\
 &= \mathbb{E}\left[\tilde{X}\tilde{X}^\top\right] - (\mathbb{E}\tilde{X})(\mathbb{E}\tilde{X})^\top - (\mathbb{E}\tilde{X})(\mathbb{E}\tilde{X})^\top + (\mathbb{E}\tilde{X})(\mathbb{E}\tilde{X})^\top \\
 &= \mathbb{E}\left[\tilde{X}\tilde{X}^\top\right] - (\mathbb{E}\tilde{X})(\mathbb{E}\tilde{X})^\top
 \end{aligned}$$

□

Theorem 4.14 (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:*

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

Proof. Left to the reader...

□

Corollary 4.1. *For any two random variables X and Y and scalars a and b :*

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

Proof. Apply Theorem 4.14 with $n = 2$, $X_1 = X$, and $X_2 = Y$.

Or, see <https://statproofbook.github.io/P/var-lincomb.html>

□

Definition 4.11 (homoskedastic, heteroskedastic). A random variable Y is **homoskedastic** (with respect to covariates X) if the variance of Y does not vary with X :

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

Otherwise it is **heteroskedastic**.

Definition 4.12 (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 4.12).

Definition 4.13 (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_i s given the X_i s 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 4.14 (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 $P(X_1 = x_1), \dots, P(X_n = x_n)$ are all equal to some shared distribution $P(X = x)$:

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

Definition 4.15 (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 $P(Y_i = y_i | X_i = x_i)$ are all equal to the same distribution $P(Y = y | X = x)$:

$$P(Y_i = y | X_i = x) = P(Y = y | X = x)$$

Definition 4.16 (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 4.17 (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 .

4.5 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 4).

```
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)
```

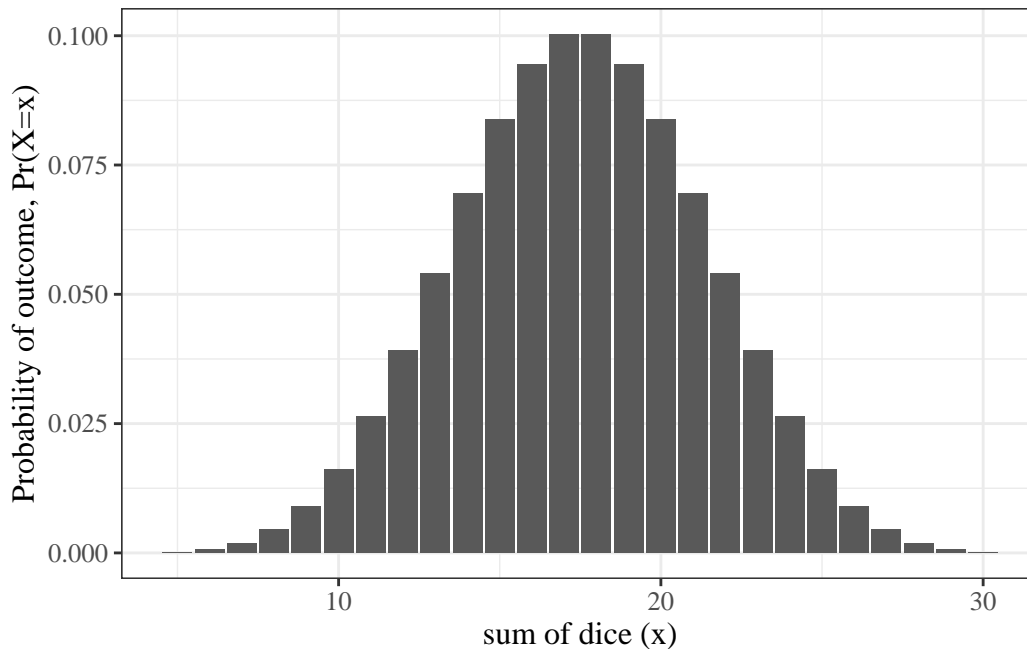


Figure 4: Distribution of the sum of five dice

In comparison, the outcome of just one die is not bell-shaped (Figure 5).

```
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)
```

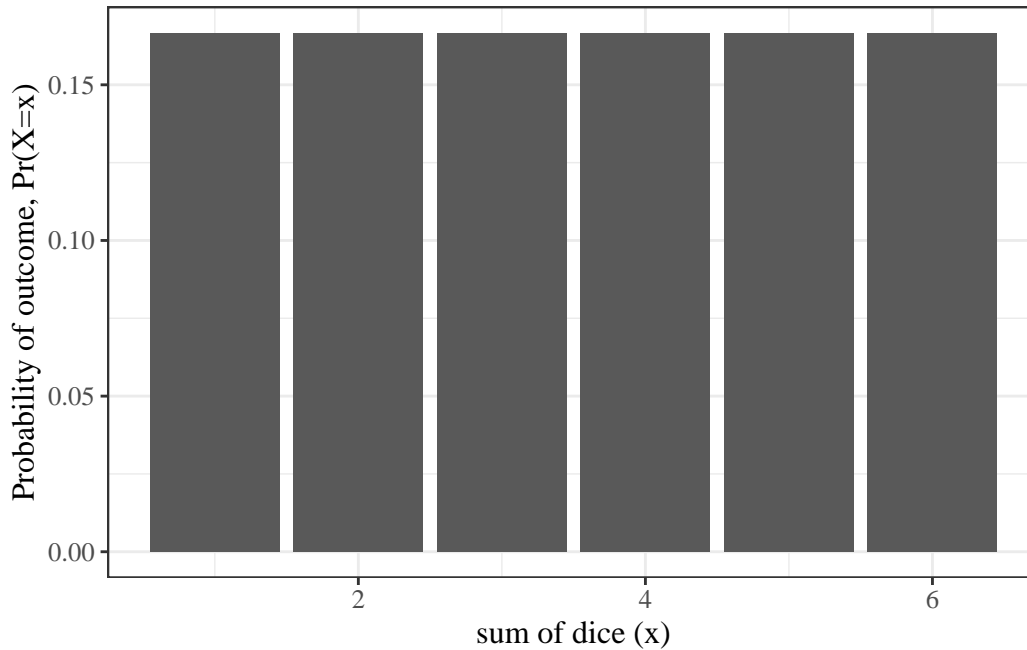


Figure 5: Distribution of the outcome of one die

What distribution does a single die have?

Answer: discrete uniform on 1:6.

5 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>.