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
library(latex2exp) # use LaTeX in R code (for figures and tables)
```

```r
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 $A_1, A_2, \ldots$ (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. *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.

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

Proof

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

Corollary 1.1.

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

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

Proof

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

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:

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

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

□

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

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

Distribution	Uses
Log-normal	Time to event outcomes

Table 2: Distributions typically used for test statistics

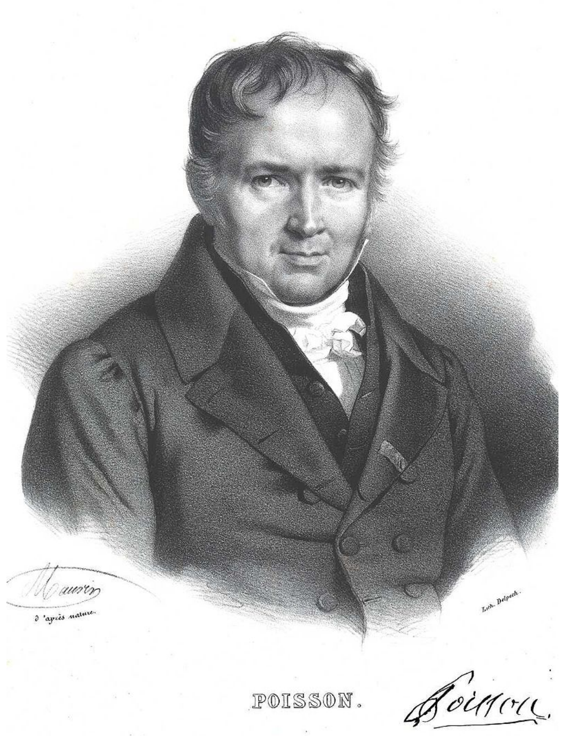
Distribution	Uses
χ^2	Regression comparisons (asymptotic), contingency table independence tests, goodness-of-fit tests
F	Gaussian model comparisons (exact)
Z (standard normal)	Proportions, means, regression coefficients (asymptotic)
T	Means, regression coefficients in Gaussian outcome models (exact)

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).

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

(see Figure 2)

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

Solution

Solution 2.2.

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

> **Theorem 2.1** (CDF of Poisson distribution)**.**
>
> $$P(Y \leq y) = e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} \tag{2}$$

(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)
```

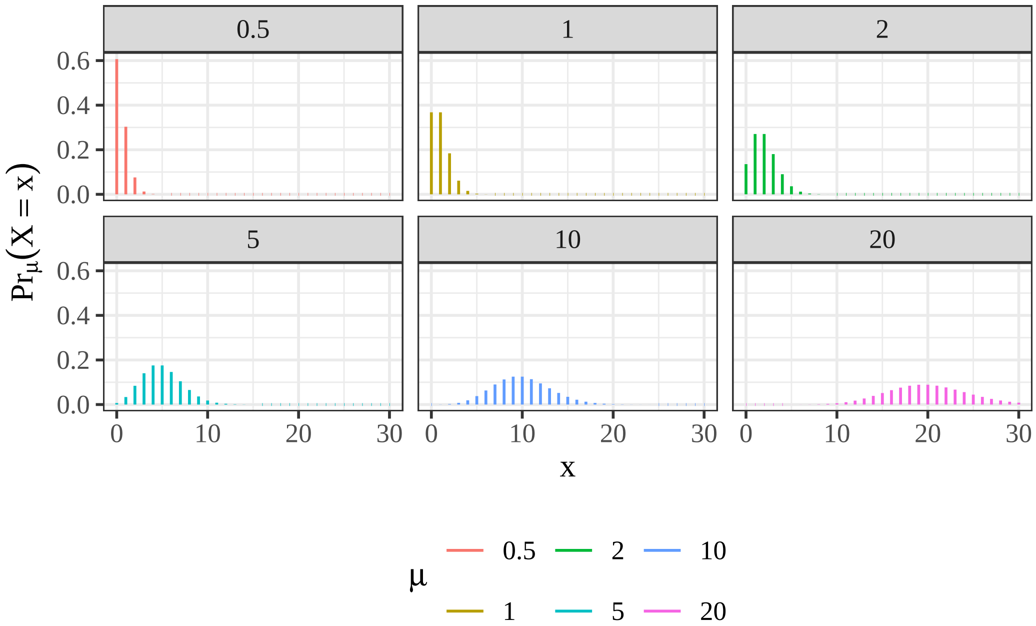


Figure 2: Poisson PMFs, by mean parameter μ

```
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)
```

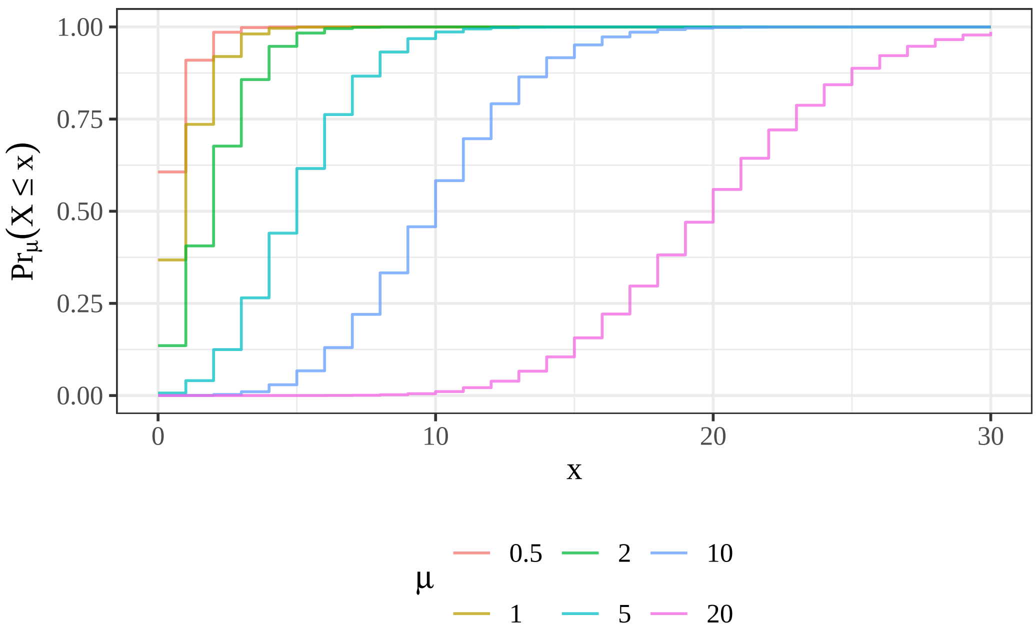


Figure 3: Poisson CDFs

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 $X \sim Pois(\mu)$, then:*

- $E[X] = \mu$
- $Var(X) = \mu$
- $P(X = x) = \frac{\mu}{x} P(X = x - 1)$
- *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 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:

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

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

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

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

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

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

$$E[Y|T=0] = 0$$

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

Theorem 2.5. *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. *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 $\mu_Z = \mu_X + \mu_Y$.*

ℹ 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. *If $Y \sim NegBin(\mu, \rho)$, then:*

- $E[Y] = \mu$

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

2.4 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.

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 $p(T = t)$ 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 $\text{h}(t)$ [2] or $\lambda(t)$, [3] is the conditional density[a] of T at t, given $T \geq t$. That is:

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

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.

[a] probability.qmd#def-pdf

Table 4: Probability distribution functions

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

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

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

3.3 Expectation

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

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

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

[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)

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:*

$$E[X] = \pi$$

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:*

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

Theorem 3.5 (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

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:

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

Example 3.1 (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:

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

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

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

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:

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

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

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

$$E[Y] = E[E[Y \mid X]]$$

*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 Fubini's theorem, 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). □

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

$$E[Y \mid Z] = E[E[Y \mid X, Z] \mid Z]$$

This is the tower rule applied conditionally on Z.

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

Example 3.2 (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 $\mathrm{E}[A_{ij}]$:

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

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

3.4 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

---

3.5 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)**.**

$$Var(X) = E[X^2] - (E[X])^2$$

Proof

Proof. By linearity of expectation, we have:

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

□

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

$$Var(Y) = E[\,Var(Y \mid X)] + Var(E[Y \mid X])$$

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

ℹ 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:

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

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:

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

□

Definition 3.11 (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.

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

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

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

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

$$Cov(X,X) = Var(X)$$

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 $\tilde{X}$,

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

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

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

Proof

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

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

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)**.**

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

ℹ Proof

Proof.

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

□

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

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

ℹ Proof

Proof. Left to the reader... □

Corollary 3.1. *For any two random variables* X *and* Y *and scalars* a *and* b*:*

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

ℹ 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:

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

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 $\mathrm{P}(X_1 = x_1), \dots, \mathrm{P}(X_n = x_n)$ 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 $X_1, \dots, X_n$.

3.6 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)
```

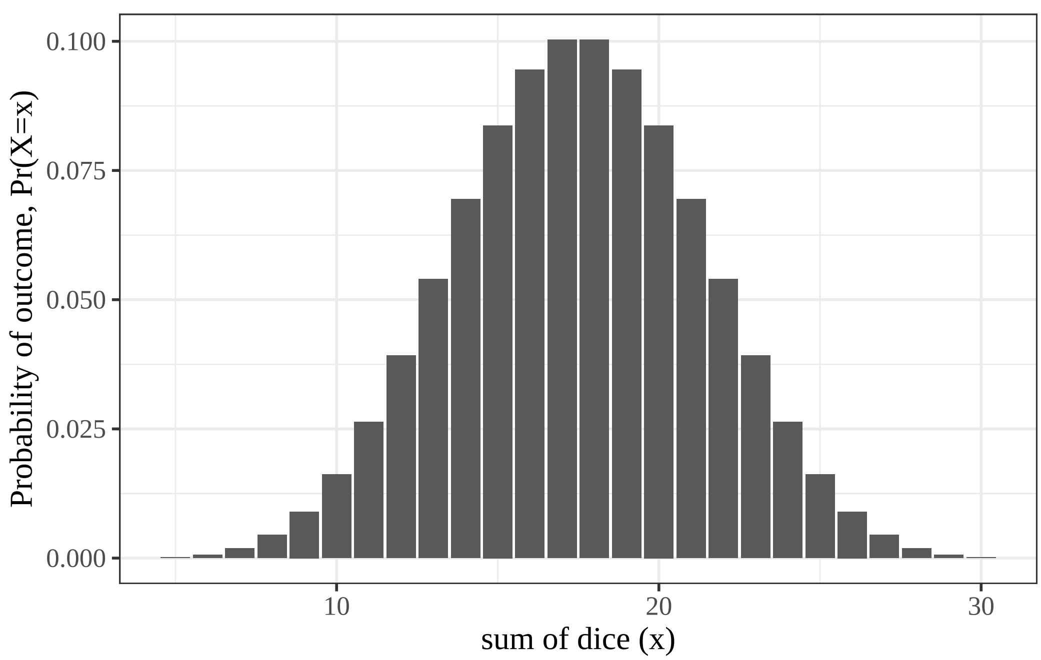


Figure 4: Distribution of the sum of five dice

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)
```

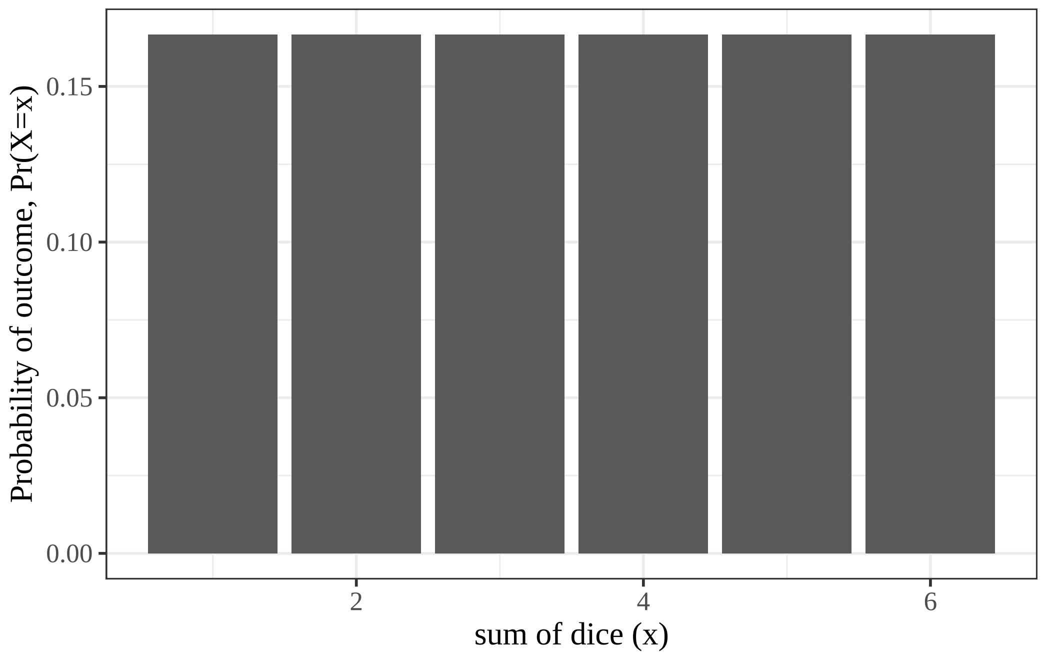


Figure 5: Distribution of the outcome of one die

What distribution does a single die have?

Answer: discrete uniform on 1:6.

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.