Statistical Theory II Practical Labs

Lab 3: Continuous Random Variables

Paulius Kazlauskas

Outline

Announcement

May 5th – Lab Exam in a computer class

Room 302

Discrete vs. Continuous Random Variables

From Discrete to Continuous

Discrete RV

- Takes countable values $y_1, y_2, \ldots$
- Described by a **PMF**: $f(y_i) = P(Y = y_i)$
- Probabilities computed by **summing**

Continuous RV

- Takes values on an interval
- Described by a **PDF**: $f(y)$
- Probabilities computed by **integrating**

Critical difference

For any continuous RV: $P(Y = c) = 0$ for any single value c.

Probability only exists over *intervals*:

$$P(a \leq Y \leq b) = \int_a^b f(y)\, dy$$

The PDF and CDF

Probability Density Function (PDF): $f(y)$

$f(y) \geq 0$ for all y, and $\int_{-\infty}^{\infty} f(y)\, dy = 1$

$f(y)$ is **not** a probability — it is a density

`d...(x)` in R evaluates $f(x)$ at a point

Cumulative Distribution Function (CDF): $F(y)$

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)\, dt$$

Always increasing, from 0 to 1

`p...(x)` in R evaluates $F(x) = P(Y \leq x)$

Computing Probabilities from the CDF

The key identity

$$P(a \le Y \le b) \;=\; F(b) - F(a) \;=\; P(Y \le b) - P(Y \le a)$$

Other useful forms

$$\begin{aligned} P(Y > a) &= 1 - F(a) \\ P(a < Y < b) &= F(b) - F(a) \\ P(Y \le a) &= F(a) \end{aligned}$$

In R — always the same pattern

```
pexp(5, 0.25) - pexp(4, 0.25)
```
$\Rightarrow P(4 < X < 5)$

```
1 - pnorm(2.11, 0, 1)
```
$\Rightarrow P(Z > 2.11)$

The d / p / q / r System

Four Functions, One System

Every distribution in R follows the same naming convention:

Prefix	What it gives	Question answered
`d`	PDF value $f(x)$	How dense is the distribution at x?
`p`	CDF value $F(x) = P(X \leq x)$	What probability lies *below* x?
`q`	Quantile: x such that $F(x) = p$	What value has probability p below it?
`r`	Random sample	Give me n draws from this distribution

Example cores: `norm`, `exp`, `unif`, `beta`, `gamma`, `binom`, `pois`

Distribution Reference Table

Distribution	Core	Key parameters	Example call
Normal	`norm`	`mean`, `sd`	`pnorm(1.96, 0, 1)`
Exponential	`exp`	`rate` $= 1/\mu$	`pexp(5, 0.25)`
Uniform	`unif`	`min`, `max`	`punif(57, 0, 60)`
Beta	`beta`	`shape1`, `shape2`	`pbeta(0.5, 10, 12)`
Gamma	`gamma`	`shape`, `rate`	`pgamma(4, 12, 4)`
Binomial	`binom`	`size`, `prob`	`dbinom(2, 3, 0.5)`
Poisson	`pois`	`lambda`	`dpois(1, 2)`

Exponential Distribution

Exponential Distribution

Setup

$X \sim \mathrm{Exp}(\lambda)$: models *waiting times*.

Rate parameter $\lambda = 1/\mu$, where μ is the mean.

PDF and CDF

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$F(x) = 1 - e^{-\lambda x}$$

Mean and Variance

$$E(X) = \frac{1}{\lambda} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Today's example

Mean $= 4$ min $\Rightarrow \lambda = 0.25$

$$P(4 < X < 5) = e^{-1} - e^{-1.25} \approx 0.0814$$

Memoryless Property of the Exponential

Statement

$$P(X \leq s + t \mid X > s) = P(X \leq t)$$

Having already waited s units gives **no information** about how much longer you will wait.

Proof sketch

$$P(X \leq s+t \mid X > s) = \frac{P(s < X \leq s + t)}{P(X > s)} = \frac{F(s+t) - F(s)}{1 - F(s)} = \frac{e^{-\lambda s} - e^{-\lambda(s+t)}}{e^{-\lambda s}} = 1 - e^{-\lambda t}$$

Why it matters

The exponential is the *only* continuous distribution with this property. It is also the continuous analogue of the geometric distribution.

Uniform Distribution

Uniform Distribution

Setup

$X \sim \mathrm{Uniform}(a, b)$: all values in $[a, b]$ equally likely.

PDF and CDF

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

$$F(x) = \frac{x-a}{b-a}$$

Mean and Variance

$$E(X) = \frac{a+b}{2}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Today's example

$X \sim U(0, 60)$:

$$P(57 \leq X \leq 60) = \frac{3}{60} = \frac{1}{20}$$

Normal Distribution

Normal Distribution

Setup

$Y \sim N(\mu, \sigma^2)$: symmetric, bell-shaped.

Note: R uses $\texttt{sd} = \sigma$, *not* variance σ^2.

PDF

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Mean and Variance

$$E(Y) = \mu \qquad \mathrm{Var}(Y) = \sigma^2$$

Standard Normal

$Z \sim N(0, 1)$: special case $\mu = 0$, $\sigma = 1$.

CDF written $\Phi(z) = P(Z \leq z)$.

Standardisation: Converting to Z

The Z-score transformation

If $Y \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

Using it to compute probabilities

$$P(Y \leq y) = P\left(Z \leq \frac{y - \mu}{\sigma}\right) = \Phi\left(\frac{y - \mu}{\sigma}\right)$$

$$P(a \leq Y \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

Today: $Y \sim N(120, 64)$, find $P(Y \leq 130)$

$$P(Y \leq 130) = \Phi\left(\frac{130 - 120}{8}\right) = \Phi(1.25) \approx 0.8944$$

Normal Probability: Useful Symmetry Rules

Symmetry of $N(0,1)$

$$P(Z \le -z) = P(Z \ge z) = 1 - \Phi(z)$$

$$P(-z \le Z \le z) = 2\Phi(z) - 1$$

Splitting around zero (for table lookups)

$$P(a \le Z \le b) = \Phi(b) - \Phi(a) \quad \text{(direct, in R always)}$$

When using printed Z-tables:

$$P(-a \le Z \le b) = P(0 \le Z \le a) + P(0 \le Z \le b)$$

In R: always use `pnorm` directly

`pnorm(1.74, 0, 1) - pnorm(-1.45, 0, 1)` — no need to split.

Percentiles and Quantiles

Definition

The p-th percentile is the value y^* such that $P(Y \leq y^*) = p$.

This is the **inverse CDF** — the `q` functions in R.

R functions

`qnorm(0.025, 0, 1)`	2.5th percentile of $N(0, 1) \approx -1.96$
`qnorm(0.975, 0, 1)`	97.5th percentile $\approx +1.96$
`qgamma(0.025, r, v)`	2.5th percentile of Gamma

Why 2.5 and 97.5?

These cut off 2.5% in each tail, leaving 95% in the middle — the basis of 95% confidence intervals you will see throughout the course.

Beta and Gamma Distributions

Beta Distribution

Setup

$Y \sim \mathrm{Beta}(\alpha, \beta)$: defined on [0, 1], useful for proportions and probabilities.

Mean and Variance

$$E(Y) = \frac{\alpha}{\alpha + \beta}$$

$$\mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Today: $Y \sim \mathrm{Beta}(10, 12)$

$$E(Y) = \frac{10}{22} \approx 0.4545$$

$$\mathrm{Var}(Y) \approx 0.0108$$

R

```
pbeta(y, shape1, shape2)
rbeta(n, shape1, shape2)
```

Gamma Distribution

Setup

$Y \sim \mathrm{Gamma}(r, v)$: models waiting times for r events at rate v.

R uses $\texttt{shape} = r$, $\texttt{rate} = v$.

Mean and Variance

$$E(Y) = \frac{r}{v} \qquad \mathrm{Var}(Y) = \frac{r}{v^2}$$

Today: $Y \sim \mathrm{Gamma}(12, 4)$

$$E(Y) = \frac{12}{4} = 3$$

$$\mathrm{Var}(Y) = \frac{12}{16} = 0.75$$

R

```
pgamma(y, shape, rate)
qgamma(p, shape, rate)
```

Normal Approximation

Normal Approximation to Other Distributions

The idea

When the exact distribution is awkward to work with (or when you only have a Z-table), you can approximate Y with a normal distribution that has the same mean and variance:

$$Y \approx N(E(Y), \operatorname{Var}(Y))$$

then standardise and use Φ.

Steps

1. Compute $\mu = E(Y)$ and $\sigma^2 = \operatorname{Var}(Y)$ from the distribution's formulas.
2. Standardise: $Z = (Y - \mu)/\sigma$.
3. Look up the resulting Z-value in the normal table (or use `pnorm`).

Normal Approximation: Two Examples

Beta(10, 12) — $P(Y > 5)$

$\mu = 0.4545$, $\sigma = \sqrt{0.0108} \approx 0.104$

$$Z = \frac{5 - 0.4545}{0.104} \approx 43.7$$

$P(Z > 43.7) \approx 0$

Makes sense: Beta is bounded in $[0, 1]$, so $P(Y > 5) = 0$ exactly.

Gamma(12, 4) — $P(Y \leq 4)$

$\mu = 3$, $\sigma = \sqrt{0.75} \approx 0.866$

$$Z = \frac{4 - 3}{0.866} \approx 1.155$$

$P(Z \leq 1.155) \approx P(Z \leq 1.15) = 0.8749$

Exact via `pgamma`: compare both!

Always compare approximation to exact

The normal approximation is better when the distribution is roughly symmetric and the tails are not extreme.

Plotting Distributions in R

Simulating and Plotting

```r
# Simulate 1000 draws and overlay the true density
y <- rnorm(1000, mean = 120, sd = sqrt(64))

# freq = FALSE converts y-axis to density (not counts)
hist(y, main = "Density", freq = FALSE)

# Add the true density curve on top
x <- seq(90, 160, 0.5)
curve(dnorm(x, 120, sqrt(64)), add = TRUE, col = "red")
```

Key arguments

`freq = FALSE`	Plot density on y-axis, not counts
`add = TRUE`	Add to existing plot rather than new window
`curve(...)`	Plot any function of `x` over a range

R Reference

R: Probability Calculations

```
# --- Exponential: lambda = 0.25 ---
pexp(5, 0.25) - pexp(4, 0.25)         # P(4 < X < 5)
1 - pexp(5, 0.25)                     # P(X > 5)

# --- Uniform: X ~ U(0, 60) ---
punif(60, 0, 60) - punif(57, 0, 60) # P(57 <= X <= 60)

# --- Normal: Y ~ N(120, 64), sd = sqrt(64) = 8 ---
pnorm(130, 120, 8)                    # P(Y <= 130)
pnorm(127, 120, 8) - pnorm(114, 120, 8) # P(114 <= Y <= 127)
pnorm(1.52, 0, 1) - pnorm(0, 0, 1) # P(0 <= Z <= 1.52)
1 - pnorm(2.11, 0, 1)                # P(Z >= 2.11)
qnorm(0.025, 0, 1)                   # 2.5th percentile
qnorm(0.975, 0, 1)                   # 97.5th percentile
```

R: Beta and Gamma

```
# --- Beta(10, 12) ---
a <- 10; b <- 12
Ey   <- a / (a + b)
Vary <- a * b / ((a + b)^2 * (a + b + 1))

Y    <- rbeta(1000, a, b)           # simulate
1 - pbeta(0.5, a, b)               # P(Y > 0.5) exact
1 - pnorm(0.5, Ey, sqrt(Vary))     # normal approx

# --- Gamma(12, 4) ---
r <- 12; v <- 4
Ey   <- r / v
Vary <- r / v^2

pgamma(4, r, v)                    # P(Y <= 4) exact
pnorm(4, Ey, sqrt(Vary))           # normal approx
```

Summary

Key Concepts

$P(X = c) = 0$ for continuous RVs

Probabilities are areas under the PDF

$P(a \leq X \leq b) = F(b) - F(a)$

$Z = (Y - \mu)/\sigma$ standardises any normal

Memoryless property: exponential only

Normal approx: match μ and σ^2, then standardise

Formulas

Dist.	$E(Y)$	$\mathrm{Var}(Y)$
$\mathrm{Exp}(\lambda)$	$1/\lambda$	$1/\lambda^2$
$U(a, b)$	$(a+b)/2$	$(b-a)^2/12$
$N(\mu, \sigma^2)$	μ	σ^2
$\mathrm{Beta}(\alpha, \beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$\mathrm{Gamma}(r, v)$	r/v	r/v^2

Bonus time

Can somebody explain (attempt) what an R Markdown is?

Can somebody show how to make one?