STAT 390 A
Statistical Methods in Engineering and Science

Week 4 Lectures – Part 1 – Spring 2023

Continuous Random Variables and Probability Distributions

Alexander Giessing

Department of Statistics

University of Washington

April 17, 2023

Outline

Recall: Examples of Continuous Random Variables

- *Spatial data. Let X be the current temperature at a random location (defined by latitude and longitude).*
  - sample space $\Omega = [0; 180] \times [0; 360]$.
  - $X(\omega) =$ current temperature at location ω.
- *Income of a randomly selected taxpayer.*
  - sample space $\Omega = [-\infty, \infty]$.
  - Let $X(\omega) = \omega$
- *Amount of precipitation per year at some location in Seattle.*
  - sample space $\Omega = [0, \infty]$.
  - Let $X(\omega) =$ inches of rainfall at location in Seattle ω

Probability Density Function (PDF)

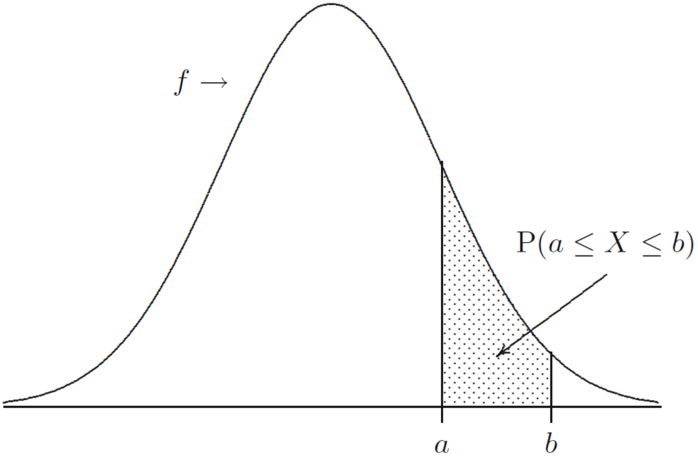
Probability Density Function

A random variable X is continuous if there exists a function $f : \mathbb{R} \to [0, \infty]$ such that for all real numbers $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

In particular, the function f satisfies $\int_{-\infty}^{\infty} f(x)dx = 1$ and $f(x) \geq 0$ for all $x \in \mathbb{R}$. We call f the probability density function of X.

Probability Density Function (Cont.)



Probability Density Function (Cont.)

- The probablity that X lies in the interval $[a, b]$ is equal to the area under the pdf of X over the interval $[a, b]$. So,

$$P(X = a) = \lim_{\varepsilon \downarrow 0} P(a - \varepsilon \leq X \leq a + \varepsilon) = \lim_{\varepsilon \downarrow 0} \int_{a-\varepsilon}^{a+\varepsilon} f(x)dx = 0.$$

  Therefore,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b) = P(a \leq X < b).$$

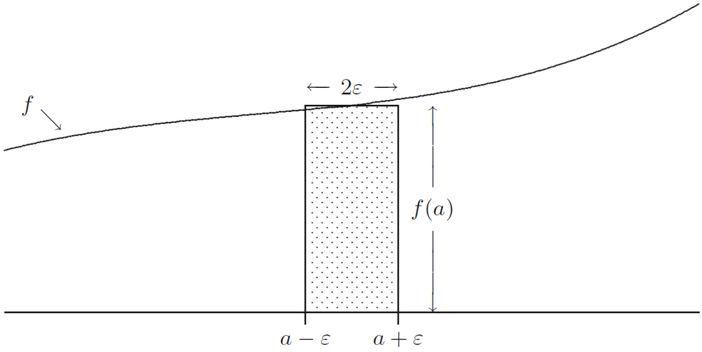
- What does $f(a)$ represent? For small $\varepsilon > 0$,

$$P(a - \varepsilon \leq X \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)dx \approx 2\varepsilon f(a).$$

  Hence, $f(a)$ can be interpreted as a (relative) measure of how likely it is tht X will take a value near a.

  Caveat: $f(a)$ is not a probability; $f(a)$ can be arbitrarily large.

Probability Density Function (Cont.)



Cumulative Distribution Function (CDF)

CUMULATIVE DISTRIBUTION FUNCTION (CONTINUOUS RANDOM VARIABLE)

Let X be a continuous random variable. The cumulative distribution function of X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(a) := P(X \leq a) = \int_{-\infty}^{a} f(x)dx \quad \text{for} \quad -\infty < a < \infty.$$

$F(a)$ is the probability that the observed value of X is at most a.

- Discrete RV have a pmf but no df, whereas continuous RV have a df but no pmf. However, discrete and continuous RV both have a cdf.
- The properties of the cdf for discrete RV also hold for the cdf of a continuous RV.
- **Facts from calculus:** $f(x) = \frac{d}{dx}F(x)$ and $P(a \leq X \leq b) = F(b) - F(a)$.

Example: Throwing Darts at Random

We want to construct a probability model for an experiment that can be described as "an object hits a disc of radius r in a completely arbitrary way". Suppose X is the distance between the hitting point and the center of the disc.

1. Since distances are nonnegative, $F(a) = P(X \leq a) = 0$ for all $a < 0$.

2. Since the object hits the disc for sure, $F(a) = 1$ for all $a > r$.

3. That the object hits the disc in a completely arbitrary way, may be interpeted as that the probability of hitting any region is proportional to the area of that region. Therefore,

$$F(a) = P(X \leq a) = \frac{\pi a^2}{\pi r^2} = \frac{a^2}{r^2} \quad \text{for } 0 \leq a \leq r.$$

4. The pdf of X is given by

$$f(x) = \frac{d}{dx} F(x) = \begin{cases} \frac{2x}{r^2} & 0 \leq x \leq r, \\ 0 & o/w. \end{cases}$$

Expectation and Change-of-Variable Formula

Expectation (Continuous Random Variable)

The expectation of a continuous random variable X is defined as

$$\mathrm{E}[X] := \int_{-\infty}^{\infty} x f(x) dx.$$

Change-of-Variable Formula (Continuous Random Variable)

Let X be a continuous random variable. For any function $g : \mathbb{R} \to \mathbb{R}$ we have,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

- **Linearity of Expectation:** $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$ for all $a, b \in \mathbb{R}$.
- **Variance:** $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

Change-of-Units Tansformation

Change-of-Units Tansformation (Continuous Random Variable)

Let X be a continuous random variable with distribution function F_X and probbility density f_X. If we change units to $Y = rX + s$ for real numbers $r > 0$ and $s \in \mathbb{R}$, then

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right) \quad \text{and} \quad f_Y(y) = \frac{1}{r} f_X\left(\frac{y-s}{r}\right) \quad -\infty < x < \infty.$$

Derivation:

$$F_Y(y) = P(Y \leq y) = P(rX + s \leq y) = P\left(X \leq \frac{y-s}{r}\right) = F_X\left(\frac{y-s}{r}\right).$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-s}{r}\right) \overset{(a)}{=} \frac{1}{r} f_X\left(\frac{y-s}{r}\right),$$

where (a) holds by the chain rule.

Quantiles/ Percentiles of Distributions

Quantiles/ Percentiles of Distributions

Let X be a continuous random variable and let $0 \leq p \leq 1$. The pth quantile or $100o$th percentile of the distribution of X is the smallest number q_p such that

$$F(q_p) = P(X \leq q_p) = p.$$

The median of a distribution is its 50th percentile.

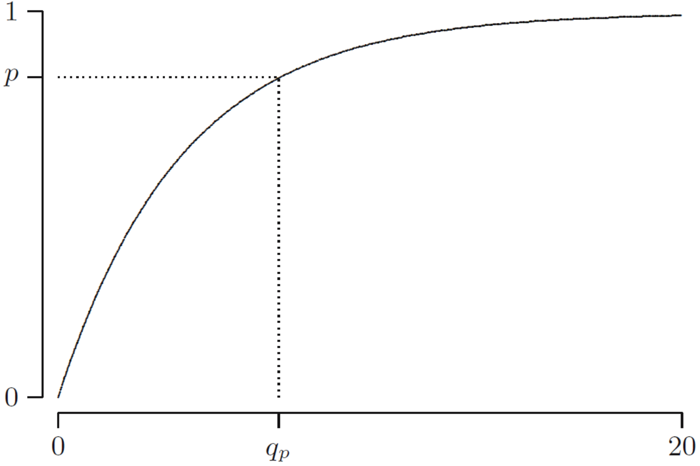
- For continuous RVs the value q_p is often easy to determine. If the cdf F is strictly increasing from 0 to 1, then

  $$q_p = F^{-1}(p),$$

  where F^{-1} is the inverse of F.

Quantiles/ Percentiles of Distributions (Cont.)

(Sketch of a strictly increasing cdf and quantile q_p).



Outline

Gaussian/ Normal Random Variable

Gaussian/ Normal Random Variable

A continuous random variable has a Gaussian (or normal) distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if its probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We denote this distribution by $N(\mu, \sigma^2)$.

- We say X follows a standard normal random variable if $X \sim N(0, 1)$.
- We write Φ for the cdf and ϕ for the df of $N(0, 1)$.

Gaussian/ Normal Random Variables (Cont.)

Derivation $\mathrm{E}[X] = \mu$.

$$\begin{aligned}
\mathrm{E}[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\overset{(a)}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (u\sigma + \mu) e^{-\frac{u^2}{2}} du \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du \\
&\overset{(b)}{=} \sigma \times 0 + \mu \times 1 \\
&= \mu,
\end{aligned}$$

where (a) follows from a change of variables $u = (x-\mu)/\sigma$ *and (b) holds because the first integral is over symmetric function (hence* $\equiv 0$*) and second integral is over the density (hence* $\equiv 1$.

Gaussian/ Normal Random Variables (Cont.)

Derivation $\mathrm{Var}(X) = \sigma^2$.

$$
\begin{aligned}
\mathrm{E}[X^2] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\overset{(a)}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (u\sigma + \mu)^2 e^{-\frac{u^2}{2}} du \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du + \mu^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du + \frac{2\sigma\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du \\
&\overset{(b)}{=} \sigma^2 \times 1 + \mu^2 \times 1 + 2\sigma\mu \times 0, \\
&= \sigma^2 + \mu^2,
\end{aligned}
$$

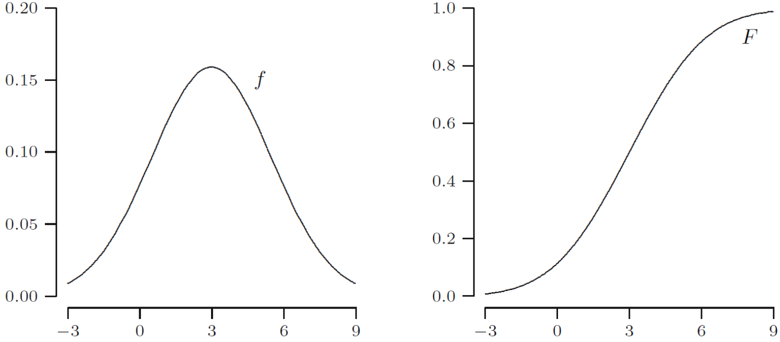
where (a) follows from a change of variables $u = (x - \mu)/\sigma$ and (b) holds by integration by parts and the previous arguments used to compute $\mathrm{E}[X]$. (For more details see your notes from the lecture).

Now compute

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

Gaussian/ Normal Random Variables (Cont.)

(Plots of df and cdf of $N(3, 6.25)$*.*



- How does the shape of the normal distribution change with mean μ and standard deviation σ? (see R-code/ illustration in class)

R-code: Plotting DF and CDF of Normal Distribution

```
> # sequence of numbers between -10 and 10 with increments of 0.005
> x <- seq(-10, 10, by = .005)
>
> # mean and sd (experiment with different values!)
> mean <- 0
> sd <- 2
>
> # df and cdf
> df <- dnorm(x, mean, sd)
> cdf <- pnorm(x, mean, sd)
>
> # plot titles
> title.df <- paste0("Density of N(", mean, ",", sd,")")
> title.cdf <- paste0("Density of N(", mean, ",", sd,")")
>
> # side-by-side plots of df and cdf
> par(mfrow = c(1,2), cex=0.8)
> plot(x,df, type = "l", xlab = "x", ylab = "density", main = title.df)
> plot(x,cdf, type = "l", xlab = "x", ylab = "density", main = title.cdf)
```

Gaussian/ Normal Random Variables (Cont.)

NORMAL RANDOM VARIABLES UNDER CHANGE OF UNITS

Let $X \sim N(\mu, \sigma^2)$. For any $r \neq 0$ and any $s \in \mathbb{R}$,

$$rX + s \sim N(r\mu + s, r^2\sigma^2).$$

- **Standardization:** If $X \sim N(\mu, \sigma^2)$, then

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

  transforms X into **standard units (SU)**.

- **Use of SUs:** How many SDs σ is the realization x above or below the mean μ?

- If $X \sim N(\mu, \sigma^2)$, then

  $$F_X(a) = P(X \leq a) = P(\sigma Z + \mu \leq a) = P\left(Z \leq \frac{a - \mu}{\sigma}\right) \equiv \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Examples: area under the normal curve

Let $X \sim N(\mu, \sigma^2)$*. What is* $P(a \le X \le b)$ *in terms of standard units?*

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

If $X \sim N(3, 2^2)$*, compute* $P(1 \le X \le 6)$*.*

$$\begin{aligned} P(1 \le X \le 6) &= \Phi\left(\frac{6-3}{2}\right) - \Phi\left(\frac{1-3}{2}\right) \\ &= \Phi(1.5) - \Phi(-1) = 0.9332 - 0.1587 = 77.45\%. \end{aligned}$$

If $X \sim N(-5, 6^2)$*, find* $P(X \ge -2)$*.*

$$\begin{aligned} P(X \ge -2) &= \Phi(\infty) - \Phi\left(\frac{-2+5}{6}\right) \\ &= 1 - \Phi(0.5) = 1 - 0.6915 = 30.85\% \end{aligned}$$

Example with R-Code: “Normal” Passengers

Suppose that the arrival time of a passenger to a train station is normally distributed with mean 11:55am and SD 10 minutes. If the train is scheduled to leave at 12:03pm, what is the chance that the passenger will catch the train?

- Let X be the arrival time relative to 12:00pm. Then, $X \sim N(-5, 10^2)$. Hence, the probability of catching the train is

$$P(X < 3) = \Phi\left(\frac{3+5}{10}\right) - 0 = 78.81\%.$$

- The probability can becomputed using R intwo ways:

```
> pnorm(0.8, mean=0, sd = 1)
[1] 0.7881446
> pnorm(3, mean -5, sd = 10)
[1] 0.7881446
```

Normal Approximation of Data Histograms

Normal approximation of data histograms

If a data histogram is bell-shaped with sample average $\bar{x}$ and sample standard deviation s, then we may approximate the distribution of the data with $N(\bar{x}, s^2)$.

"Infinitely" many applications:

- Human physical characteristics (height, weight)
- Measurement error in scientific experiments
- Scores of exams (this is what we mean by "curving" the grades!)
- Nearly any real-valued measurement whose exact distribution is unknown
- ...

Reason: Averages of random variables are approximately normal distributed.

Example: Normal Approximation of Blood Pressure

Suppose that the blood pressures of the users aged 25 to 34 in a drug study averaged out to 121mm with an SD of 12.5mm. The data histogram follows a normal curve/ is bell shaped.

- Approximately, what percentage of users have blood pressure between 96mm and 133.5mm?

  Let $X \sim N(121, 12.5^2)$. The percentage is given by:

$$P(96 \leq X \leq 133.5) = \Phi\left(\frac{133.5 - 121}{12.5}\right) - \Phi\left(\frac{96 - 121}{12.5}\right)$$
$$= 0.8413 - 0.0228 = 81.85\%.$$

- If the top 10% of users need to be singled out for further investigation, what would be the cut-off point for their blood pressure?
  - Convert the percentile into the standard units (SU): upper 10th percentile = 90th percentile = $z_{0.90} = 1.28$ SU.
    ```
    > qnorm(0.9, mean = 0, sd = 1)
    [1] 1.281552
    ```
  - Convert the SU into the original unit:

$$z_{0.90} \times \sigma + \mu = 1.28 \times 12.5mm + 121mm = 137mm.$$

Example: Normal Approximation of IQ

The IQ in a particular population (as measured by a standard test) is known to be approximately normally distributed with mean 100 and SD 15.

- What percentage of the population has an IQ at least 125?

  Let $X \sim N(100, 15^2)$. The percentage is given by:

  $$P(X \geq 125) = 1 - \Phi\left(\frac{125 - 100}{15}\right) = 4.78\%.$$

- What is the 95th percentile of the population?

  - Convert the percentile into the standard units (SU):
    95th percentile $= z_{0.95} = 1.64$ SU

    ```
    > qnorm(0.95, mean = 0, sd = 1)
    [1] 1.644854
    ```

  - Convert the SU into the original unit:

    $$z_{0.95} \times \sigma + \mu = 1.64 \times 15 + 100 = 124.6.$$

Outline

Uniform Distribution

Uniform Distribution

A continuous random variable has a uniform distribution on the interval $[\alpha, \beta]$ if its probability density function f is given by

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha \leq x \leq \beta, \\ 0 & o/w. \end{cases}$$

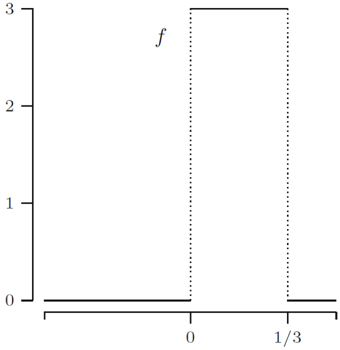
We denote this distribution by $Unif(\alpha, \beta)$.

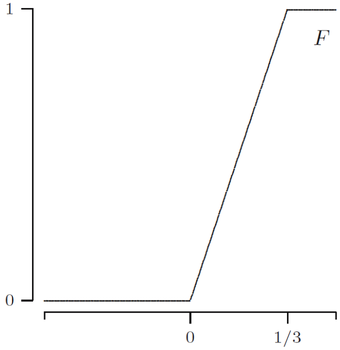
Recall the example of throwing darts at random:

- We assumed that the probability of hitting any area on the disc was proportional to the area of that disc.
- This means that the dart hits are uniformly distributed over the disc. This is a 2-dimensional uniform distribution defined on the area of a circle with radius r! Can you generalize this to a 3-dimensional sphere?
- Note: The distance X from the center to where the dart hits the disc is not uniformly distributed.

Uniform Distribution (Cont.)

(Plots of df and cdf of $Unif(0, 1/3)$*.)*





Exponential Distribution

Exponential Distribution

A continuous random variable has an exponential distribution with parameter $\lambda > 0$ if its probability density function f is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & o/w. \end{cases}$$

We denote this distribution by $Exp(\lambda)$.

- The $Exp(\lambda)$ can be thought of as continuous version of $Geo(p)$, which models waiting times. The parameter $\lambda > 0$ is the "rate" at which "successes" occur.

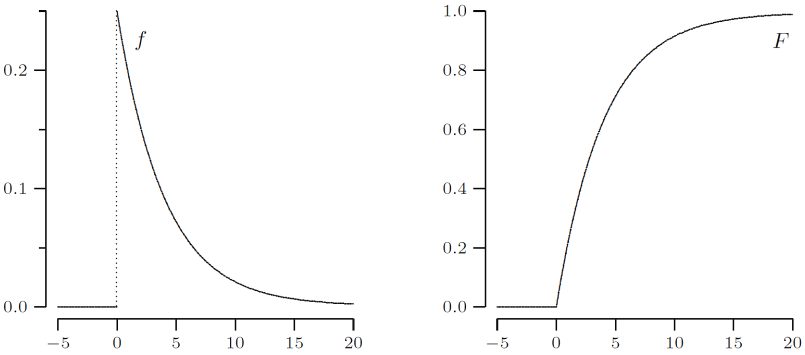
- **Memoryless property:** If $X \sim Exp(\lambda)$, then for all $s, t > 0$,

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

(Derivation.)

Exponential Distribution (Cont.)

(Plots of df and cdf of $Exp(0.25)$*.)*



Pareto Distribution

Pareto Distribution

A continuous random variable has a Pareto distribution with parameter $\alpha > 0$ if its probability density function f is given by

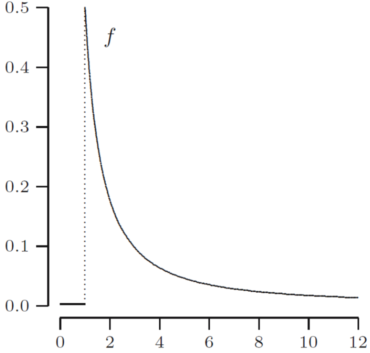
$$f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{if } x \geq 1, \\ 0 & o/w. \end{cases}$$

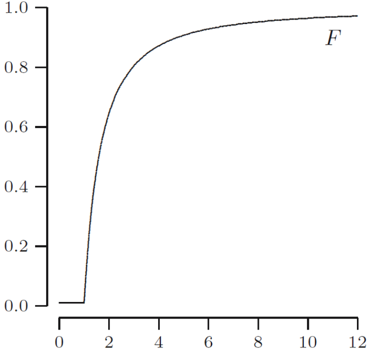
We denote this distribution by $Par(\alpha)$.

- Used to model wealth distibutions in societies, city sizes, earthquake rupture areas, size of insurance claims, size of commerical companies.
- Motivated by empirical observations (economist Vifredo Pareto).

Pareto Distribution (Cont.)

(Plots of df and cdf of $Par(0.5)$.)





Quantile-Quantile-Plots (QQ-Plots)

How can we tell different distributions apart?

Quantile-Quantile-Plot

- Purpose: To check whether the data $x_1, \ldots, x_n$ follow a distribution F.
- Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be the order statistics.
- The ith order statistic $x_{(i)}$ satisfies

$$x_{(i)} \approx \left(\frac{i - 0.5}{n}\right) \text{th (empirical) percentile of the data.}$$

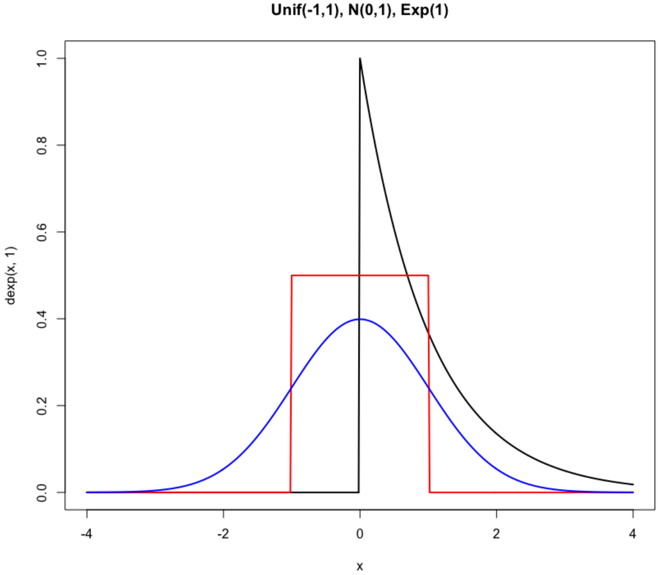
- The Q-Q plot shows the pairs

$$\left(\underbrace{F^{-1}\left(\frac{i-0.5}{n}\right)}_{\text{theoretical quantiles of } F}, \underbrace{x_{(i)}}_{\text{quantiles of the data}} \right), \quad i = 1, \ldots, n.$$

- If the pairs fall approximately on a straight line, conclude that the data follows distribution F.

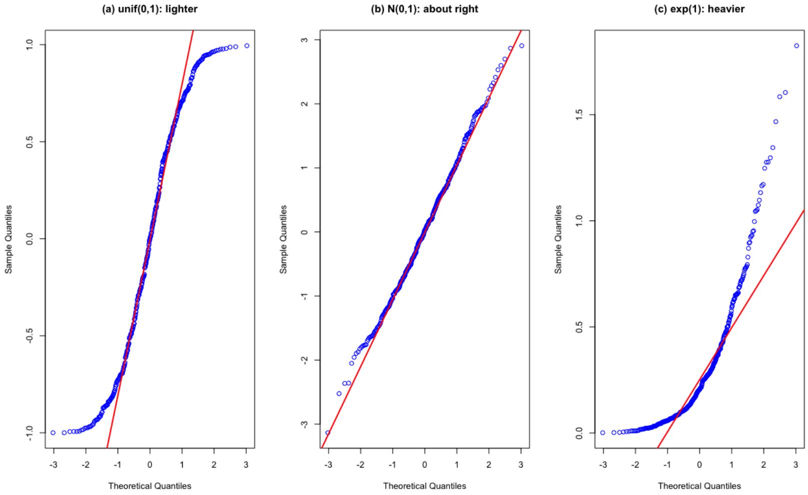
Example with R-Code: QQ-Plots

Comparing three different distributions via QQ-plots.



Example with R-Code: QQ-Plots (Cont.)

We generate data from $Unif(0,1)$, $N(0,1)$, $Exp(1)$.



Example with R-Code: QQ-Plots (Cont.)

```
> ## Generate first plot
> x = seq(-4, 4, 0.02)
> # exponential density
> plot(x,dexp(x,1),col="black",type="l",lwd=2)
> # uniform density
> lines(x,dunif(x,-1,1),col="red",type="l",lwd=2)
> # normal density
> lines(x,dnorm(x), col="blue", lwd=2)
> title("Unif(-1,1), N(0,1), Exp(1)")
>
> ## Generate second plot
> par(mfrow = c(1,3), cex = 0.8)
> x <- runif(400,-1,1) # Uniform random numbers
> qqnorm(x, col="blue", main="")
> qqline(x, lwd=2, col="red") # add a line connecting Q_1 and Q3
> title("(a) unif(0,1): lighter")
> x <- rnorm(400) #Normal random numbers
> qqnorm(x, col="blue", main=""); qqline(x, lwd=2, col="red")
> title("(b) N(0,1): about right")
> x <- rexp(400, 3) #exponential
> qqnorm(x, col="blue", main=""); qqline(x, lwd=2, col="red")
> title("(c) exp(1): heavier")
```