Background for Lesson 5

1 Cumulative Distribution Function

The cumulative distribution function (CDF) exists for every distribution. We define it as $F(x) = P(X \leq x)$ for random variable X. If X is discrete-valued, then the CDF is computed with summation $F(x) = \sum_{t=-\infty}^{x} f(t)$ where $f(t) = P(X = t)$ is the probability mass function (PMF) that we have already seen. If X is continuous, the CDF is computed with an integral $F(x) = \int_{-\infty}^{x} f(t)dt$ where $f(t)$ is the probability density function (PDF).

Example: Suppose $X \sim \text{Binomial}(5, 0.6)$. Then $F(1) = P(X \leq 1) = \sum_{-\infty}^{1} f(t) = \sum_{t=-\infty}^{-1} 0 + \sum_{t=0}^{1} \binom{5}{t} 0.6^t (1-0.6)^{5-t} = \binom{5}{0} 0.6^0 (1-0.6)^{5-0} + \binom{5}{1} 0.6^1 (1-0.6)^{5-1} = (0.4)^5 + 5(0.6)(0.4)^4 \approx 0.087$.

Example: Suppose $Y \sim \text{Exp}(1)$. Then $F(2) = P(Y \leq 2) = \int_{-\infty}^{2} e^{-t} I_{\{t \geq 0\}} dt = \int_{0}^{2} e^{-t} dt = -e^{-t}|_0^2 = -(e^{-2} - e^0) = 1 - e^{-2} \approx 0.865$.

The CDF is convenient for calculating probabilities of intervals. Let a and b be any real numbers with $a < b$. Then the probability that X falls between a and b is equal to $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$. This concept is illustrated in Figure 1.

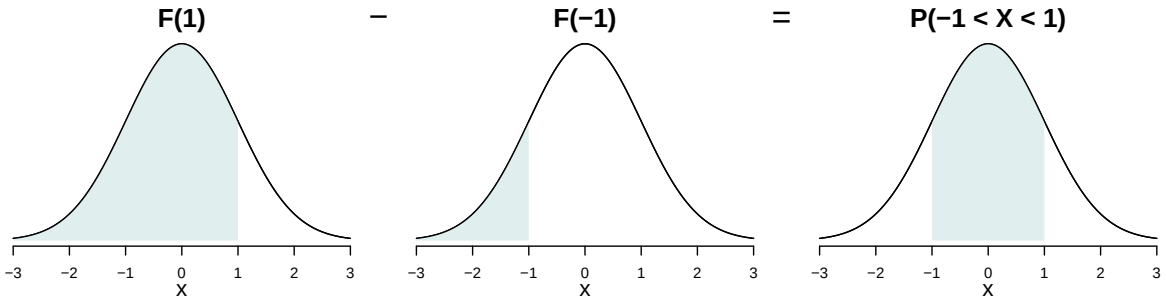

Figure 1: Illustration of using the CDF to calculate the probability of an interval for continuous random variable X. Probability values are represented with shaded regions in the graphs.

2 Quantile Function

The CDF takes a value for a random variable and returns a probability. Suppose instead that we start with a number between 0 and 1, call it p, and we wish to find the value x so that $P(X \leq x) = p$. The value x which satisfies this equation is called the p quantile (or $100p$ percentile) of the distribution of X.

Example: In a standardized test, the 97th percentile of scores among all test-takers is 23. Then 23 is the score you must achieve on the test in order to score higher than 97% of all test-takers. We could equivalently call $q = 23$ the .97 quantile of the distribution of test scores.

Example: The middle 50% of probability mass for a continuous random variable is found between the .25 and .75 quantiles of its distribution. If $Z \sim \text{N}(0, 1)$, then the .25 quantile is -0.674 and the .75 quantile is 0.674. Therefore, $P(-0.674 < Z < 0.674) = 0.5$.

3 Probability Distributions in R

Each of the distributions introduced in Lesson 3 have convenient functions in R which allow you to evaluate the PDF/PMF, CDF, and quantile functions, as well as generate random samples from the distribution. To illustrate, Table 1 lists these functions for the normal distribution.

Table 1: R functions for evaluating the normal distribution $\text{N}(\mu, \sigma^2)$.

Function	What it does
`dnorm(x, mean, sd)`	Evaluate the PDF at `x` (`mean` $= \mu$ and `sd` $= \sqrt{\sigma^2}$).
`pnorm(q, mean, sd)`	Evaluate the CDF at `q`.
`qnorm(p, mean, sd)`	Evaluate the quantile function at `p`.
`rnorm(n, mean, sd)`	Generate `n` pseudo-random samples from the normal distribution.

These four functions exist for each distribution, where the `d...` function evaluates the density/mass, `p...` evaluates the CDF, `q...` evaluates the quantile, and `r...` generates a sample. Table 2 lists the `d...` functions for some of the most popular distributions. The `d`

can be replaced with `p`, `q`, or `r` for any of the distributions, depending on what you want to calculate. For more details, enter `?dnorm` to view R's documentation page for the normal distribution. As usual, replace the `norm` with any distribution to read the documentation for that distribution.

Table 2: R functions for evaluating the density/mass function for several distributions.

Distribution	Function	Parameters
Binomial(n, p)	`dbinom(x, size, prob)`	`size` $= n$, `prob` $= p$
Poisson(λ)	`dpois(x, lambda)`	`lambda` $= \lambda$
Exp(λ)	`dexp(x, rate)`	`rate` $= \lambda$
Gamma(α, β)	`dgamma(x, shape, rate)`	`shape` $= \alpha$, `rate` $= \beta$
Uniform(a, b)	`dunif(x, min, max)`	`min` $= a$, `max` $= b$
Beta(α, β)	`dbeta(x, shape1, shape2)`	`shape1` $= \alpha$, `shape2` $= \beta$
N(μ, σ^2)	`dnorm(x, mean, sd)`	`mean` $= \mu$, `sd` $= \sqrt{\sigma^2}$
t_ν	`dt(x, df)`	`df` $= \nu$

Example: Suppose $X \sim$ Binomial$(5, 0.6)$. Then we can evaluate $F(1) = P(X \leq 1) \approx 0.087$ in R with `pbinom(q=1, size=5, prob=0.6)`. Note also that `qbinom(p=0.087, size=5, prob=0.6)` will return 1 as expected.

Example: Suppose $Y \sim$ Exp(1). The middle 80% of probability mass is located between the 0.1 and 0.9 quantiles. To find these quantiles of the Exp(1) distribution, save them as a vector in R: `a = c(0.1, 0.9)` followed by `qexp(p=a, rate=1)` which returns the vector $(0.105, 2.303)$. Therefore, we have $P(0.105 < Y \leq 2.303) = 0.8$.

Practice: The remaining lessons require many calculations with distributions. Here are a few problems to practice in R. Answers are given in blue.

1. Let $X \sim$ Pois(3). Find $P(X = 1)$. (0.149)
2. Let $X \sim$ Pois(3). Find $P(X \leq 1)$. (0.199)
3. Let $X \sim$ Pois(3). Find $P(X > 1)$. (0.801)
4. Let $Y \sim$ Gamma$(2, 1/3)$. Find $P(0.5 < Y < 1.5)$. (0.078)

5. Let $Z \sim \mathrm{N}(0,1)$. Find z such that $P(Z < z) = 0.975$. (1.96)

6. Let $Z \sim \mathrm{N}(0,1)$. Find $P(-1.96 < Z < 1.96)$. (0.95)

7. Let $Z \sim \mathrm{N}(0,1)$. Find z such that $P(-z < Z < z) = 0.90$. (1.64)

4 Probability Distributions in Excel

Excel also provides convenient functions for evaluating probability distributions. There are two primary functions which we can modify to accomplish each of the four tasks of computing a PDF/PMF, computing a CDF, computing a quantile, and generating a pseudo-random sample. These functions are demonstrated for the normal distribution in Table 3.

Table 3: Excel functions for evaluating the normal distribution $\mathrm{N}(\mu, \sigma^2)$. Replace `x`, `mean` $= \mu$, `standard_dev` $= \sqrt{\sigma^2}$, and `probability` with numbers or cell references.

Function	What it does
`NORM.DIST(x, mean, standard_dev, FALSE)`	Evaluate the PDF at `x` (`cumulative = FALSE`).
`NORM.DIST(x, mean, standard_dev, TRUE)`	Evaluate the CDF at `x` (`cumulative = TRUE`).
`NORM.INV(probability, mean, standard_dev)`	Evaluate the quantile function at `probability`.
`NORM.INV(RAND(), mean, standard_dev)`	Generate one sample (`probability=RAND()`).

The `RAND()` function generates a pseudo-random draw from the Uniform$(0,1)$ distribution, which if passed through the normal quantile function, produces a draw from the normal distribution.

The `.DIST` and `.INV` functions are available for most of the distributions listed in Table 4, which gives the PDF/PMF function for each. More information can be obtained by searching these functions in the Excel help menu.

Example: Suppose $X \sim \text{Binomial}(5, 0.6)$. Then we can evaluate $F(1) = P(X \leq 1) \approx 0.087$ in Excel by entering `= BINOM.DIST(1, 5, 0.6, TRUE)`. Note also that `= BINOM.INV(5, 0.6, 0.087)` (where `trials` $= 5$, `probability_s` $= 0.6$, and `alpha` $= 0.087$) will return 1 as expected.

Example: Suppose $Y \sim \text{Exp}(3)$. The middle 80% of probability mass is located between the 0.1 and 0.9 quantiles. To find these quantiles of the Exp(1) distribution, save 0.1 and

Table 4: Excel functions for evaluating the density/mass function for several distributions. Replace the arguments with numbers or cell references.

Distribution	Function	Parameters
Binomial(n, p)	`BINOM.DIST(x, trials, probability_s, FALSE)`	`trials` $= n$, `probability_s` $= p$
Poisson(λ)	`POISSON.DIST(x, mean, FALSE)`	`mean` $= \lambda$
Exp(λ)	`EXPON.DIST(x, lambda, FALSE)`	`lambda` $= \lambda$
Gamma(a, b)	`GAMMA.DIST(x, alpha, beta, FALSE)`	`alpha` $= a$, `beta` $= 1/b$
Beta(α, β)	`BETA.DIST(x, alpha, beta, FALSE)`	`alpha` $= \alpha$, `beta` $= \beta$
N(μ, σ^2)	`NORM.DIST(x, mean, standard_dev, FALSE)`	`mean` $= \mu$, `standard_dev` $= \sqrt{\sigma^2}$
t_ν	`T.DIST(x, deg_freedom, FALSE)`	`deg_freedom` $= \nu$

0.9 in two cells, say `A1` and `A2`. Note that Excel does not have a `EXPON.INV()` function, so we rely on the fact that the the exponential distribution is a special case of the gamma distribution with $a = 1$. We calculate the quantiles by entering `= GAMMA.INV(A1, 1, 1/3)` and `= GAMMA.INV(A2, 1, 1/3)` which yields 0.035 and 0.768 (note that the `GAMMA` functions in Excel use a scale parameter instead of a rate parameter, hence we use 1/3 instead of 3). Therefore, we have $P(0.035 < Y \leq 0.768) = 0.8$.

Practice: Now try the practice exercises from Section 3 using Excel.