Lecture 4: Sampling and Probability Distributions

Thomas Chadefaux

PO7001: Quantitative Methods I

Probability Theory

Basic terminology

- Sample space: collection of all possible outcomes of an experiment
  - E.g., rolling a die: $S = \{1, 2, 3, 4, 5, 6\}$
- An event is a subset of the sample space
  - E.g., the event A that an even number is obtained is $A = \{2, 4, 6\}$.

Probability Axioms

- The probability of any event is non-negative

$$P(A) \geq 0 \; \forall A \in S, \text{ where S is the sample space}$$

- The probability of 'anything' occurring among all possible events is 1

$$P(S = 1)$$

- If two events are disjoint, the probability that one or the other will occur is the sum of their individual probabilities.

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Random Variables

- A variable: characteristics, quantity that can be measured or counted
- A random variable: variable that can take a set of values with some associated probability.
  - More formally: a random variable is a real-valued function that is defined on S, the sample space. I.e., a function that maps the sample space S to the real numbers.
    - E.g., number of heads in an experiment with three coin tosses:

$$X(\{H, T, H\}) = 2$$
$$X(\{H, H, H\}) = 3$$
$$X(\{T, T, H\}) = 1$$

    - E.g., drawing a random person from a population and recording her height. This height is a random variable (in this case the sample space is the list of all possible people, and the function is a mapping of the form:

$$f(i) = \begin{cases} 178 & \text{if } i = John \\ 164 & \text{if } i = Beth \\ 192 & \text{if } i = Ella \\ \vdots & \vdots \end{cases}$$

Continuous and discrete R.V.

- If a variable can only take on a finite number of values, we call it discrete.
  - e.g., number of students in the classroom
- A variable that can take on any real value is called continuous
  - e.g., temperature

Distribution of Random Variables

- A random variable has a probability distribution, which specifies the probability that its value takes on a certain value or falls within any given interval.
- Random variables can be
  - discrete $\rightarrow$ probability *mass* function
  - continuous $\rightarrow$ probability density function

Probability Distributions

- Analogous to a frequency distribution, but derived from theory rather than observed data
- A probability distribution is a mathematical description of a random process in terms of the probabilities of its events.
- A probability distribution is like a frequency distribution where $N = \infty$.
- Note: terminology can be quite confusing, as many names refer to the same thing, with small (and inconsistent) variations. For practical purposes:
  - Probability density function = probability distribution function = probability function = probability distribution
  - However, be careful not to confuse "density function" with *cumulative* density function (more below)

Probability (mass) functions for discrete variables

- For a discrete variable, the probability function is

$$f(x) = P(X = x)$$

  - E.g., for a coin toss,

$$f(x) = \begin{cases} 0.5 & \text{if } \ x = \mathit{Heads} \\ 0.5 & \text{if } \ x = \mathit{Tails} \end{cases}$$

- Probability distribution for a single coin flip:

Event	Probability
Heads	0.5
Tails	0.5

- Simple example: the discrete uniform distribution has probability mass function $f(x) = \frac{1}{N}$

Probability (density) functions for continuous variables

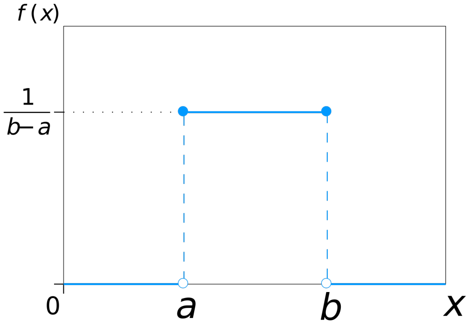
- Just like in the discrete case, the probability distribution function ('p.d.f.') assigns a probability to every possible outcome.
- More precisely, the pdf f is such that the probability that X takes a value in the interval is the integral of f over the interval:

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

E.g., the Uniform Distribution

For example, the uniform distribution has pdf:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



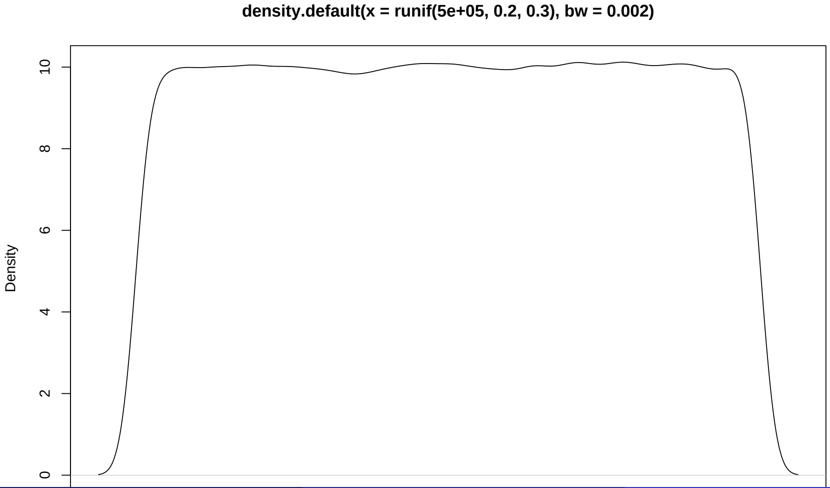
E.g., the Uniform Distribution

- Note that the pdf can take values greater than 0
  - Exercise: in R, plot the density function of a uniform distribution over the interval $[0.2, 0.3]$

E.g., the Uniform Distribution

- Note that the pdf can take values greater than 0
  - Exercise: in R, plot the density function of a uniform distribution over the interval $[0.2, 0.3]$

```
plot(density(runif(500000, 0.2,0.3), bw=0.002))
```



The *Cumulative* distribution function

The cumulative distribution function, $F(x)$, (note the use of a capital letter) is the function

$$F(x) = P(X \leq x) \text{ for } -\infty < x < \infty$$

For example, suppose you flip a coin twice times and observe the number of Heads. Then

$$F(-1) = 0$$

$$F(0) = \frac{1}{4}$$

$$F(0.5) = \frac{1}{4}$$

$$F(1) = \frac{3}{4}$$

$$F(2) = 1$$

$$F(23) = 1$$

Note: law of large numbers

- Law of large numbers: as N increases, the sample mean of a distribution approaches its theoretical mean
- e.g.:

```
mean(rnorm(10))
```

```
## [1] -0.02142639
```

```
mean(rnorm(1000))
```

```
## [1] 0.05273785
```

```
mean(rnorm(100000))
```

```
## [1] 0.001950097
```

```
mean(rnorm(1000000))
```

```
## [1] 0.0006181567
```

Important aside: Notation

- Greek vs Roman letters
  - greek letters used for *population* statistics: μ is the mean, σ is the standard deviation, σ^2 is the variance, etc.
  - Roman letters for *sample* statistics: $\bar{x}$, s and s^2

Important Distributions

The Bernoulli distribution

- A Bernoulli random variable has only two possible values: 0 and 1.
- A Bernoulli(p) random variable is defined by $P(X = 1) = p$
  - For example, the toss of one fair coin follows a Bernoulli(.5) distribution.
- The Bernoulli(p) random variable has probability mass function over possible outcomes k:

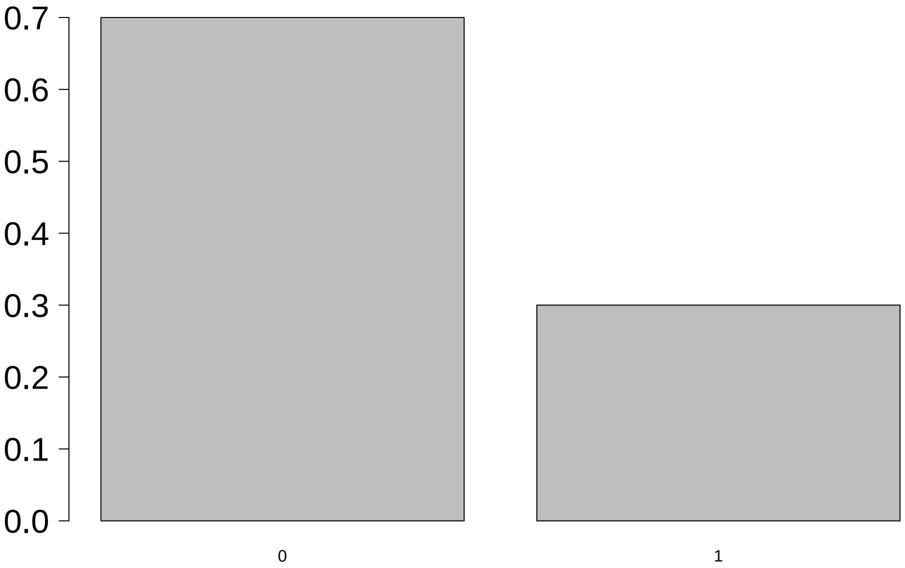
$$f(k, p) = \begin{cases} p & \textit{if } \text{k=1} \\ 1 - p & \textit{if } \text{k=0} \end{cases}$$

- For Bernoulli(p),

$$\mu = p, \sigma^2 = p(1 - p)$$

  - Exercise: prove that for Bernoulli(2p), $\sigma^2 = 2p(1 - 2p)$
- The Bernoulli distribution is a special case of the binomial distribution, i.e., the case where $n = 1$ (see below)
  - Note: we often talk about a Bernoulli 'trial' to denote an experiment where the result is 1 or 0 (success/failure), the trials are independent, and the probability of success stays the same (p)
- Exercise: graph the probability mass function of a Bernoulli(.3) variable.

PMF of A Bernoulli(.3) variable



Binomial Distribution

- Counts the number of successes in Bernoulli trials
- E.g., let X = number of Heads from flipping a (fair) coin 5 times
  - Note: like all variables, this takes a particular outcome and converts it into a number
  - This variable could take any number in [0,5]. i.e., we could get 0 Heads, 1 Heads, ... 5 Heads
- Question: What is the probability that we get 2 Heads, 3 Heads, etc?
- Possible outcomes: $\{H, T, H, T, T\}$, $\{T, T, T, H, T\}$, etc.
  - How many possible outcomes? 2 for the first flip, 2 for the second, etc. So $2^5 = 32$ equally likely possibilities.
- First, let's think about the probability that we get 0 Heads. There is only one way to get this: $\{T, T, T, T, T\}$, which is one of the 32 possible outcomes. So $P(X = 0) = \frac{1}{32}$
- How about $P(X = 1)$? There are 5 ways to get that, so $P(X = 1) = \frac{5}{32}$
- The binomial distribution counts the number of successes in Bernoulli trials, and there is a formula for it, though we won't go into details

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial Distribution: an example

- E.g., what is the probability that we get one Heads if we toss a coin twice?

$$P(X = k) = \binom{2}{1} .5^1 (1 - .5)^{2-1} = \frac{2 \times 1}{1(2-1)!} \times .25 = 0.5$$

Check in R:

```r
dbinom(1, 2, .5)
```

```
## [1] 0.5
```

Binomial Distribution: another example

- E.g., what is the probability that we get 12 Heads if we toss a coin 20 times?

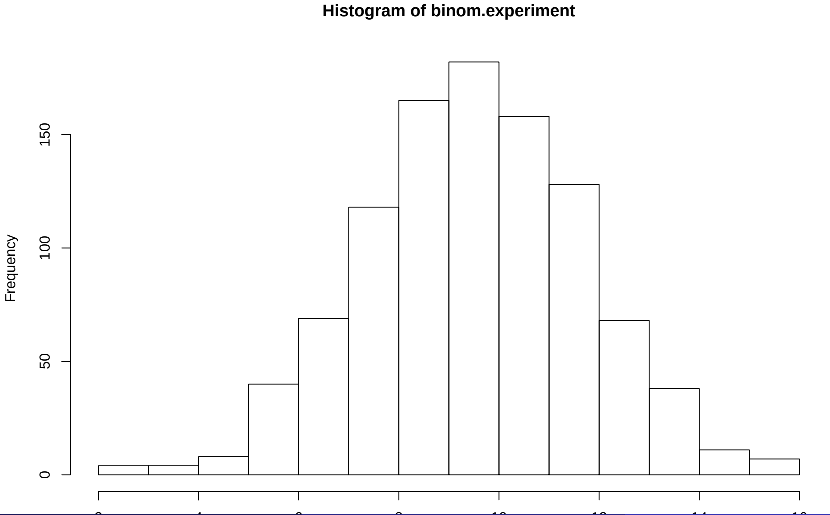
$$
\begin{aligned}
P(X = k) &= \binom{20}{12} .5^12(1 - .5)^{20-12} && (1)\\
&= \frac{20 \times 19 \times \ldots \times 2 \times 1}{12!(20-12)!} \times 0.0000009536742 && (2)\\
&\approx 0.12 && (3)
\end{aligned}
$$

```r
dbinom(12, 20, .5)
```

```
## [1] 0.1201344
```

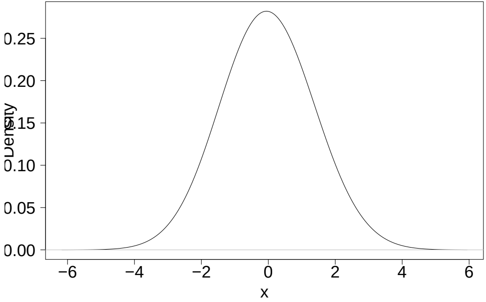
Generate a random Binomial Distribution

```
# Let our experiment be the flipping of a coin 20 times. We repeat that exp
binom.experiment <- (rbinom(1000, 20, .5))
hist(binom.experiment)
```



The normal distribution

- The most important distribution in statistics
- It is:
  - Symmetric
  - continuous
  - unimodal
  - follows a specific mathematical form involving two parameters: the mean and the variance. We write for example $X\tilde{}N(\mu, \sigma^2)$ to say that X is normally distributed with mean μ and variance σ^2.
- A variable X is normally distributed if, loosely, it has a 'bell' like curve.



Normal curve: p.d.f.

- It has the pdf

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

- With this formula, we can actually reconstruct the curve. For example, let $X \tilde{} N(0,1)$. Then:

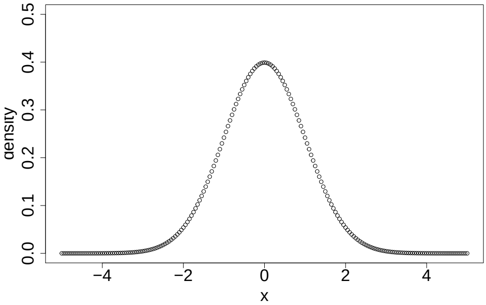
$$f(0) = \frac{1}{1\sqrt{2\pi}} e^{0} =$$

```r
# Check with R
dnorm(0)
```

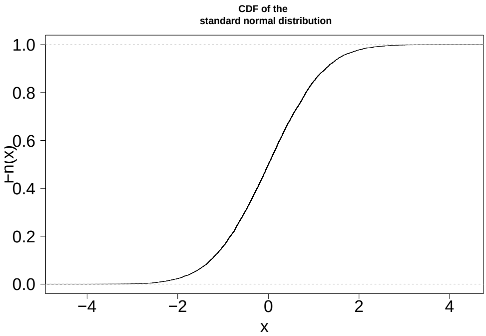
```
## [1] 0.3989423
```

Normal curve: p.d.f.

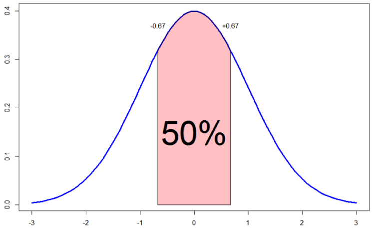
```
plot(c(-5,5), c(0,0.5), type='n', xlab='x', ylab='density', cex.axis=2, cex
for(i in seq(-5,5, 0.05)){
points(i, dnorm(i))
}
```

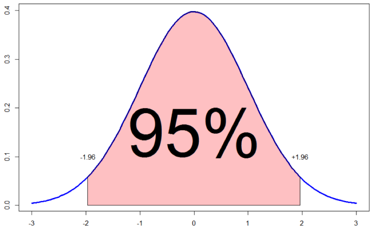


Normal curve: p.d.f.



The area under the normal curve





Standardized scores (z-scores)

You often need to standardize a variable that follows a normal distribution, in particular to transform it into a *standard* normal distribution.

$$z = \frac{x - \mu}{\sigma}$$

Many other distributions that we will come across

- Exponential distribution
- lognormal distribution
- t-distribution
- F-distribution
- χ^2 (chi-squared) distribution

Probability distributions in R

- d, p, q, and r functions
  - d: You know x and want the *density* at this point. I.e., f(x)

```
r  dnorm(0)
## [1] 0.3989423
```

  - p: you know x and want the *area* up to this point. I.e., it gives you F(x)

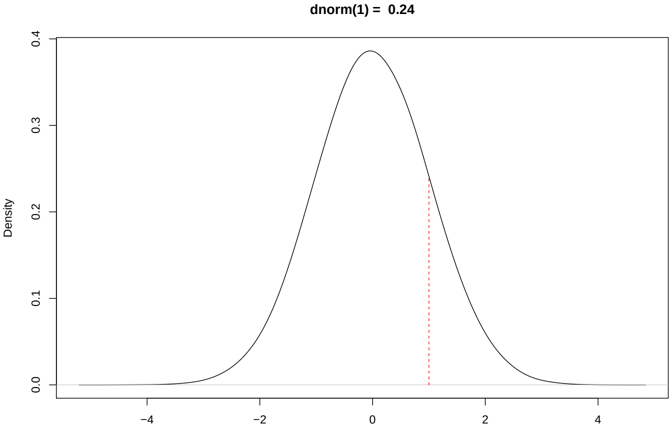
```
r  pnorm(0)
## [1] 0.5
```

  - q: you have the area and want to know the corresponding x. I.e., it gives you the inverse of the cdf. For example, you want to know which point splits the area in two:

```
r  qnorm(0.5)
## [1] 0
```

  - r: you want to generate random numbers from that distribution
- These functions work on many distributions:
  - rnorm() for the normal distribution
  - rchisq() for the chi-squared distribution
  - rf() for the F-distribution
  - rbinom() for the binomial
  - rt() for the t distribution

Examples



- For a large number of samples, the distribution *of the sample means* will be normally distributed, *no matter what the shape of the original distribution*