

# Probability Distributions – A First Look

James H. Steiger

Department of Psychology and Human Development  
Vanderbilt University

- 1 Introduction
- 2 Discrete and Continuous Probability Distributions
  - Discrete Distributions
  - Continuous Distributions
  - Quantiles
- 3 Quantiles
  - Summary of General Facts
- 4 Distribution Calculations in R
  - Calculating Probability with R
  - Relative Tail Probabilities
  - Relative Tail Probabilities
- 5 *p*-Value Calculation

# Introduction

- In this module, we learn how to perform probability distribution calculations in R.
- We are presenting basic practical material with as little background information as possible.
- There is a detailed module on probability theory later in the course.
- For now, we simply want a basic understanding of how to do the calculations, why we need them, and how they are used.

# Introduction

- In this informal introduction, you may think of probability simply as the long run equivalent of relative frequency.
- Since relative frequencies are numbers between zero and one, so are probabilities.
- A discrete probability distribution for a random variable  $X$  is a function assigning a probability to a set of possible numerical outcomes for  $X$ .
- You may express the function as a table, a mathematical formula, or as a probability distribution plot.
- In characterizing distributions, it is important to distinguish between two fundamental types of random variables — *discrete* and *continuous*.

# Discrete Distributions

- A discrete random variable  $X$  takes on a countably finite number of possible values
- The discrete random variable  $X$  has a *probability distribution function*, or *pdf*, which assigns each possible outcome a probability between 0 and 1.
- The sum of the probabilities across all the possible outcomes is 1.

# Discrete Distributions

## The Probability Function

- A classic example of a discrete random variable is the random variable  $X$  representing an idealistic die throw for a perfectly fair die.
- $X$  is the random variable,  $x$  the values it can take on. The *probability function*  $p(x)$  is defined as  $\Pr(X = x)$ , i.e., the probability that the random variable  $X$  takes on the value  $x$ .
- Imagine a completely fair die, in which all the integer outcomes from 1 to 6 are possible.
- We can present it as a table.

# Discrete Distributions

## The Probability Function

$X$	$p(x)$
6	1/6
5	1/6
4	1/6
3	1/6
2	1/6
1	1/6

# Discrete Distributions

## The Probability Function

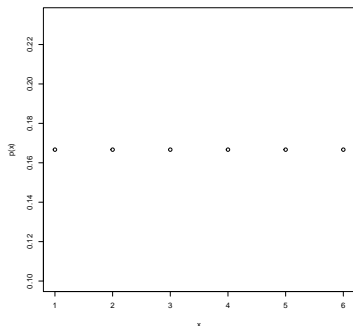
- Since the probabilities must add up to 1, a fair die with 6 outcomes must have a probability of  $1/6$  for each of the possible outcomes.
- Here is a plot.



# Discrete Distributions

## The Probability Function

```
> x <- 1:6  
> p <- rep(1/6,6)  
> plot(x,p, ylab="p(x)")
```



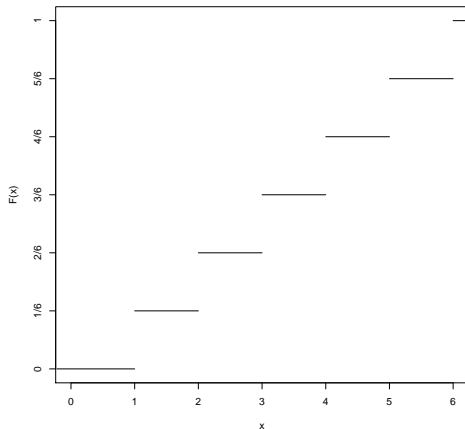
# Discrete Distributions

## The Cumulative Probability Function

- The *cumulative probability function*  $F(x)$  is defined as  $F(x) = \Pr(X \leq x)$ .
- For a fair die, what is  $F(3)$ ?  $F(4.5)$ ?  $F(17)$ ?  $F(-3)$ ?
- Note that the cumulative probability function is a step function, with steps at the values of  $X$  with non-zero probability.
- Note that it is defined at values of  $X$  that can never occur.
- Note also that the probability function  $p(x)$  is *the amount by which the cumulative probability function  $F(x)$  changes at  $x$ .*

# Discrete Distributions

## The Cumulative Probability Function



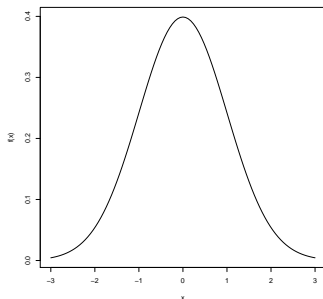
# Introduction

- Continuous distributions are used to model situations where, in principle, the number of outcomes is uncountably infinite because all values within a given range are possible.
- For example, suppose that for a random variable  $X$ , all values between 1 and 6 have equal likelihood.
- What is  $\Pr(X = 2)$ ?
- We cannot compute this probability the way we did for a fair die, because now there are infinitely many possibilities. If all of them have equal probability, there is no specifiable nonzero value we can assign to all the numbers and still have the probabilities sum to 1.
- So, for a continuous random variable  $X$ , the probability that  $X$  takes on any specific value is undefined — it is infinitesimally small.

# Introduction

- Consider the famous “standard normal curve” shown below for the random variables  $X$ .
- If we cannot define the probability that  $X = 1$ , then what is being plotted when we plot the normal curve?

```
> curve(dnorm(x), -3, 3, xlab="x", ylab="f(x)")
```



# Introduction

## Probability Density

- It is not probability that is being plotted, rather it is *probability density*.
- That is why the *Y*-axis has  $f(x)$  rather than  $p(x)$ .  
Probability density is defined as the rate that cumulative probability is increasing at  $x$ .
- Hence, probability density is the derivative of the cumulative probability function.
- Conversely, the cumulative probability function  $F(x)$  is the integral of  $f(x)$  from  $-\infty$  to  $x$ , i.e., the area under the probability density curve.
- Since probabilities must sum to 1, the total area under the probability density curve for any properly defined distribution is 1.

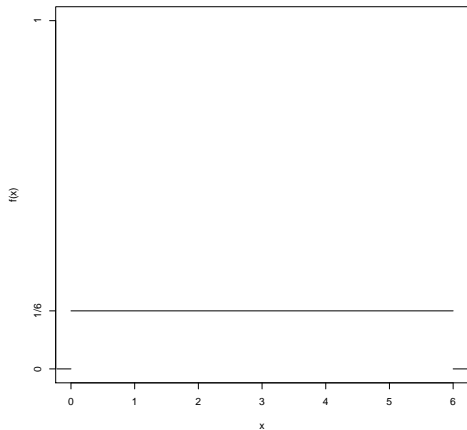
# Introduction

## Probability Density

- Consider a uniform  $(0, 6)$  random variable  $X$  that takes on all values between 0 and 6 with equal probability.
- Intuitively, what should be the probability that  $0 \leq X \leq 3$ ?
- Let's draw the pictures of  $f(x)$  and  $F(x)$ .

# Introduction

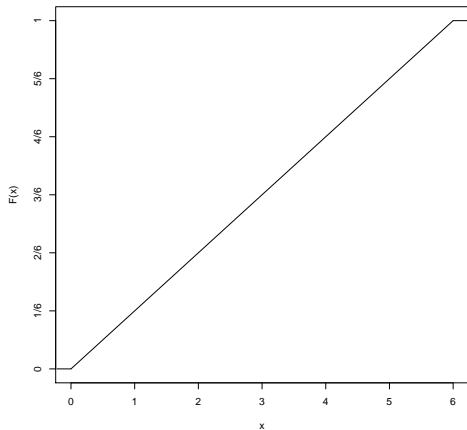
## Probability Density





# Introduction

## Probability Density



## Quantiles

- A *quantile* is that point in a distribution at or below which a certain proportion of cases fall. In other words, the  $q$ th quantile is that point  $x$  at which  $F(x) = q$ .
- For example, the .25 quantile is a point at or below which 25% of the cases fall.
- Some common examples of quantiles include quartiles, deciles, and percentiles.
- Quartiles break a distribution into 4 equal parts. Deciles break the distribution into tenths, etc. One commonly sees statements in mathematical notation such as  $Q_2 = D_5$ , i.e., the second quartile is the fifth decile.

# Quantiles

## Percentiles

- The most famous kind of quantile is the *percentile*. A percentile is that point in a distribution at or below which a certain percentage of the cases fall. For example,  $P_{50}$ , the 50<sup>th</sup> percentile, is that point at or below which half the cases fall.
- Since, in finite (discrete) distributions, quantiles are not uniquely defined, it is common to talk about the percentile value of a given score, or the score that is at a certain percentile.
- Generally the percentile value of a given score is uniquely defined, while the score that is at a given quantile may not be.
- Consider this concrete example. Four women have 0,1,6, and 8 children, respectively. What is the percentile value of the score 6 in this distribution?
- On the other hand, what score is at the 50<sup>th</sup> percentile?
- Some authors deal with this issue by defining the quantile  $q$  as the lowest point  $x$  in the distribution at or below which  $F(x) = q$ .

## Summary of General Facts

- Discrete distributions define the probability of an event, the probability of an interval, and the cumulative probability.
- Continuous distributions define probability density, the probability of an interval, and the cumulative probability.
- For discrete distributions, the probabilities of the individual outcomes must sum to 1.
- For continuous distributions, the total area under the probability density curve is 1, and the area under the density curve between any two points is the probability that  $X$  takes on a value in that interval.
- To compute the probability of an interval, we simply take the difference of the two cumulative probabilities. That is,

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

# Calculating Probability with R

- R provides a set of functions for performing probability calculations for many of the most important distributions.
- The key function types are
  - 1 *d*, for calculating probability (discrete r.v.) or probability density (continuous r.v.)
  - 2 *p*, for calculating cumulative probability.
  - 3 *q*, for calculating inverse cumulative probabilities, or *quantiles*.
- To perform these calculations, you need to know the code for the distribution you are working with. Some common codes are:
  - 1 *norm*. The normal distribution.
  - 2 *t*. Student's *t* distribution.
  - 3 *f*. The *F* distribution.
  - 4 *chisq*. The chi-square distribution.
- You also need to know the *parameters* of the distribution to fully specify it.
- You combine the function type and the distribution name to obtain the name of the function you need to perform a particular calculation.

## Calculating Probability with R

- What value of the standard normal distribution is at the 95th percentile?  

```
> qnorm(.95)
```

```
[1] 1.644854
```
- What is the probability that a *t*-variable with 19 degrees of freedom is less than or equal to 1.00?  

```
> pt(1.00,19)
```

```
[1] 0.8350616
```
- What is the probability that an observation from a normal distribution with mean 500 and standard deviation 100 will be between 600 and 700?  

```
> pnorm(700,500,100)-pnorm(600,500,100)
```

```
[1] 0.1359051
```

## Relative Tail Probabilities

- In some cases, probability calculations can yield surprising results.
- This is especially true with relative tail probability calculations.
- Suppose you have some attribute for which an individual must have a score above 160 in order to qualify for some occupation.
- Group A has a mean of 100 and a standard deviation of 15 on this attribute, while group B has a mean of 100 and a standard deviation of 17 on the attribute.
- Of those qualifying for the occupation, what percentage will be from Group B? From Group A? (Assume that Group A and Group B are equally represented in the population.)

## Relative Tail Probabilities

- Since the two groups have identical means and only slightly different standard deviations, it might seem that the representation at the tails should be fairly well balanced.
- However, let's do the calculation.

```
> GroupA <- 1 - pnorm(160,100,15)
> GroupB <- 1 - pnorm(160,100,17)
> Relative.Odds <- GroupB/GroupA
> GroupB.Probability <- Relative.Odds/(1 + Relative.Odds)
> GroupA
[1] 3.167124e-05
> GroupB
[1] 0.0002082423
> Relative.Odds
[1] 6.575122
> GroupB.Probability
[1] 0.8679889
```



## Relative Tail Probabilities

- The phenomenon is striking. On average, Group B is no better than Group A on this attribute.
- But in the upper tail, Group B is much better represented than Group A.
- Notice also that the phenomenon becomes more extreme as the cutoff point gets higher.

## *p*-Value Calculation

- In statistical hypothesis testing, it is common to test a hypothesis about a parameter with a “test statistic” that is some function of an estimate of that parameter.
- For example, we test a hypothesis that a population mean  $\mu$  is equal to 100 by examining a test statistic  $t$  that is a function of an estimate of how much  $\mu$  deviates from 100.
- If the statistical null hypothesis is true, then the test statistic has a known distribution, the “null sampling distribution.”
- If a value occurs that is unlikely to occur in that distribution, we consider the result as evidence against the null.
- In *one-sided* (or -tailed) testing situations, evidence to reject the null can come at only one end of the number line, while in *two-sided* (or -tailed) situations, the evidence can come at either end.

# *p*-Value Calculation

- Connected with this idea is a value called the *significance level* or *p*-value.
- The *p*-value is the probability of getting a result as extreme or more extreme as the observed value. If the test is one-sided, the calculation is performed in the correct direction for rejection. If the test is two-sided, the calculation is performed toward the nearest rejection side, then the resulting probability is doubled.
- With these ideas in mind, we can produce a rule for computing *p*-values that works well for continuous sampling distributions, but is more controversial for discrete distributions.
- But, in general, the lower the *p*-value, the “more significant” the result.
- In particular, if the *p*-value is less than  $\alpha$ , we say that the result is significant at the  $\alpha$  level.
- Some people have been led to believe that this means that the lower the *p*-value, the more powerful the effect shown in the result, but this is most definitely not the case.

## *p*-Value Calculation

- To see how this works, suppose you are testing the hypothesis that  $\mu_1 = \mu_2$  with a *t*-statistic with 24 degrees of freedom. The value of the statistic is 2.54. What is the *p*-value? The probability of getting a result more extreme on the positive side is  

```
> 1-pt(2.54,24)
```

```
[1] 0.00898735
```
- However, since this is a hypothesis about equality, it could be rejected with either a very low or a very high value of the test statistic. It is a two-sided test. So we must double the above probability to get a correct 2-sided *p*-value  

```
> 2 * (1-pt(2.54,24))
```

```
[1] 0.0179747
```
- Since the value is lower than 0.05, we reject the null hypothesis at the 0.05 level, but not at the 0.01 level, since the value is larger than 0.01.