Sample Statistical Workflow:
Random Variables and Distributions

Adam Davey

January 11, 2015

Distributions underlie all of statistical theory and practice. Success at biostatistics requires their mastery.

Adam

1 Preamble

You frame a research hypothesis and collect some data with which to test the hypothesis. From a frequentist perspective, we seek to determine the probability of observing these data D if the null hypothesis is true (H_0). We can translate this into a formal probability statement, such as $Pr(D|H_0)$. A Bayesian approach, which is beyond the scope of this course, attempts to more directly estimate the quantity we are usually interested in, $Pr(H_0|D)$, something that requires different assumptions. In this course, we will consider statistical models that are based around a variety of different discrete and continuous statistical distributions. It is essential that we are able to move between different characteristics of distributions to solve for quantities of interest. Our goal is to begin understanding when each kind of distribution might apply and to start developing our intuition about distributions, their shapes, and factors associated with them.

2 Random Variables

Variables are said to be "random" when their values cannot be known or determined in advance. Different values of random variables can occur with different probabilities.If the number of values that a random value can assume is *finite* and *countable*, then the distribution from which it is drawn is said to be `discrete`. If instead, the number of values is *infinite* or *uncountable*, then the variable is said to be `continuous`.

3 Characteristics of Probability Distributions

The function that links values of a function with their probabilities is called the probability distribution.

3.1 Probability Density Function

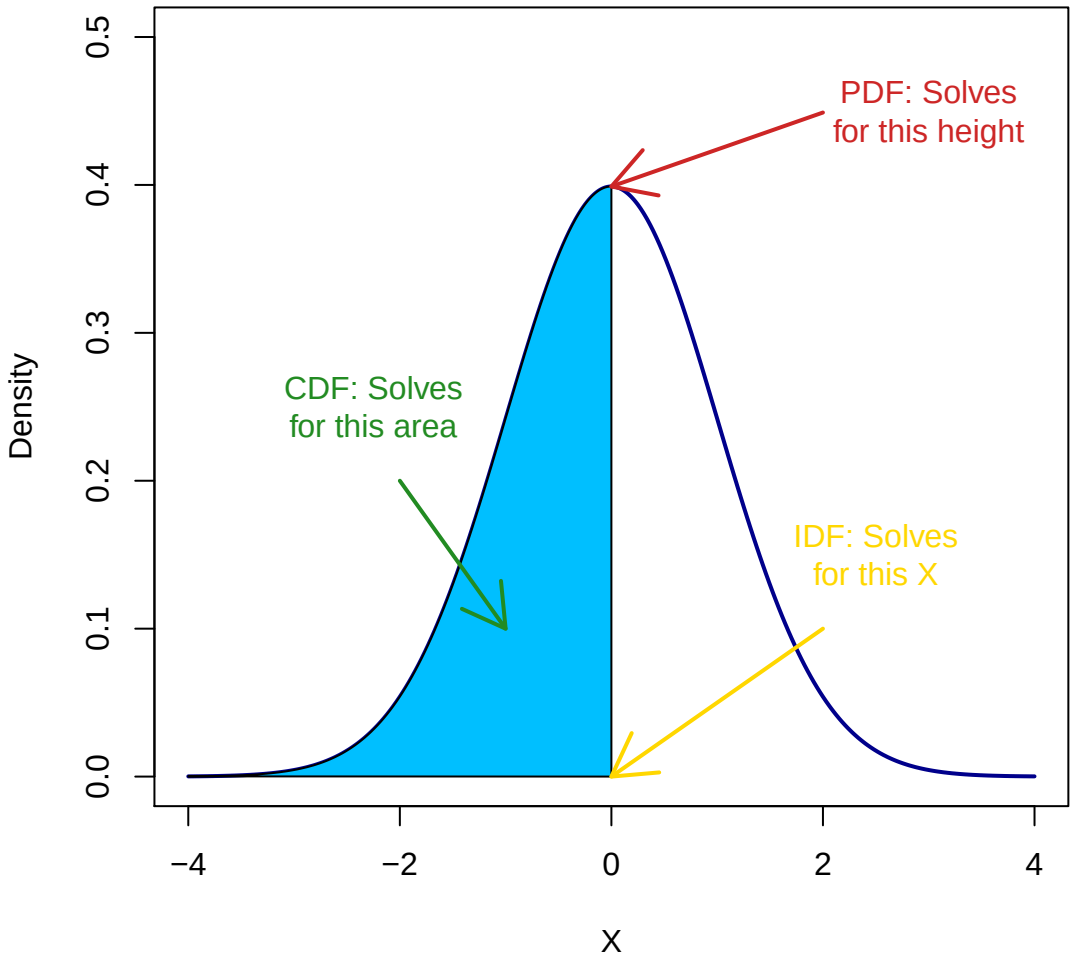
We can plot the probability density function, or PDF, which represents the relative probability of every possible value of X. Simplistically, since every experiment must have *some* outcome, the sum over all possible outcomes of an experiment must be equal to one. In other words, the area under the PDF must equal 1. With a continuous function, the probability at a single point is 0, so instead, we talk about the probability over a range of values on the random variables. (If the function is discrete, then we talk about the *probability mass function*, PMF.)

3.2 Cumulative Distribution Function

We know that summing over the entire range of possible values must be 1 if a function is a probability density function. However, it is also useful to sum over some range of values, say a value $\leq x$. This sum gives us the cumulative distribution function, or CDF. It follows that the probability for values $> x$ can be found by $1 - CDF$.

3.3 Inverse Distribution Function

Finally, it is often necessary to calculate the value of a function associated with a given cumulative probability. The function to go from a CDF to the value of a function is called the inverse distribution function, or IDF. The IDFs for many commonly used statistical distributions (e.g., normal) do not have closed form solutions and so are either approximated or built up from other distributions.
The figure below illustrates each function as it applies to a prototypical probability distribution.



4 Distributions

In this course, we will focus on the most commonly used statistical distributions. These include the uniform, Bernoulli, binomial, chi-squared, F, normal, Student's t, and Poisson distributions.

4.1 Uniform Distribution

A uniform distribution, denoted as $unif(a, b)$ assigns equal probabilities to all values between a and b. In the discrete case, we could think about the roll of a single fair die where each of the values $x \in \{1, 2, 3, 4, 5, 6\}$ are equally probable.

The PDF for a uniform distribution $unif(a, b)$ is given by $PDF = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

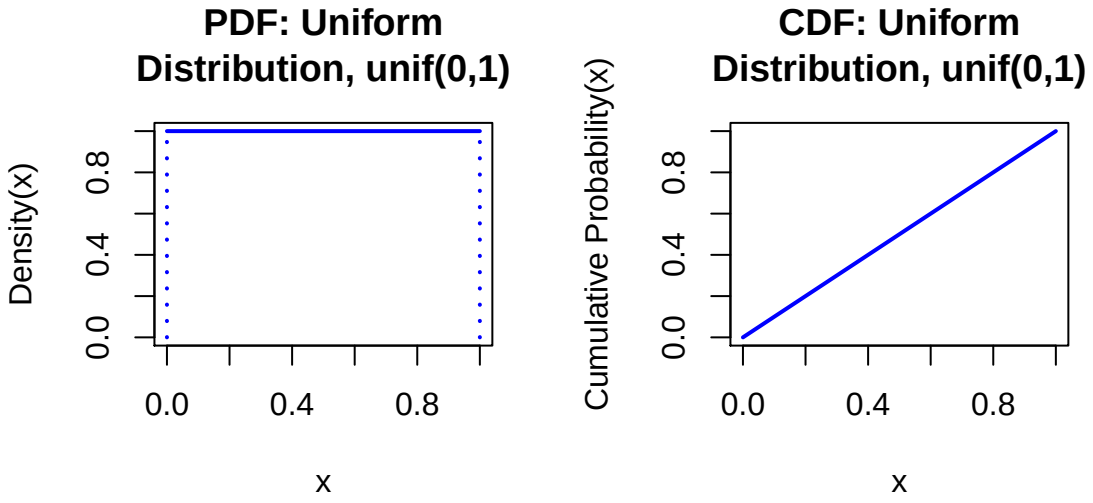
$CDF = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$

$IDF = a + p(b-a), \text{for } 0 < p < 1$

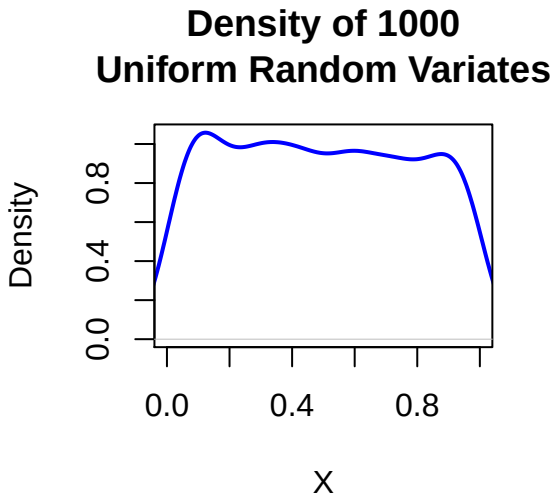
$Mean = \frac{a+b}{2}$

$Variance = \frac{(b-a)^2}{12}$

Graphically, for a continuous uniform distribution over the range of $[0, 1)$, the PDF and CDF would look like this.



In Stata, uniform random variates can be generated using `generate x = runiform()`. Because they can be obtained/calculated directly, Stata does not include facilities to calculate PDF or CDF of uniform variables. Below is a density plot based on a sample of 1000 uniform random variates.



4.2 Bernoulli

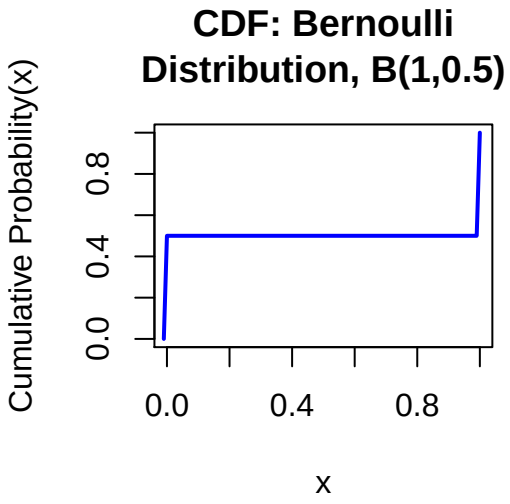
The Bernoulli distribution is a natural choice to represent a single binary event, such as a coin toss. The event has a probability of success, say coming up heads, with probability p and tails with probability $1-p$. The Bernoulli distribution is a special case of the Binomial distribution (considered below) with $n = 1$ and so is denoted by $B(1, p)$. The PMF for a Bernoulli distribution is given by

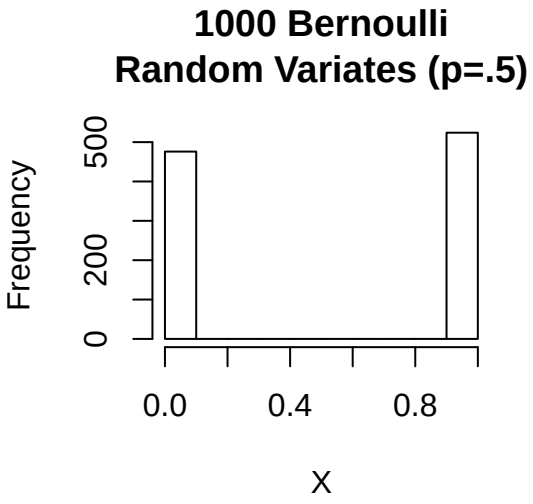
$$PMF = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

$$CDF = \begin{cases} 0 & \text{for } k < 0 \\ 1-p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$$

$$Mean = p$$

$$Variance = p(1-p)$$





In Stata, Bernoulli random variates can be generated using `generate x = rbinom(1,p)`, where p is the probability of success (i.e., 0.5 for a single fair coin toss). We consider how to obtain the PMF and CDF using Stata below when we discuss the binomial distribution.

4.3 Binomial

The binomial distribution extends a Bernoulli distribution to consider multiple simultaneous events. For example if the focus is on a set of events (i.e., what is the expected distribution of heads when repeatedly tossing 10 fair coins), then the binomial distribution applies.

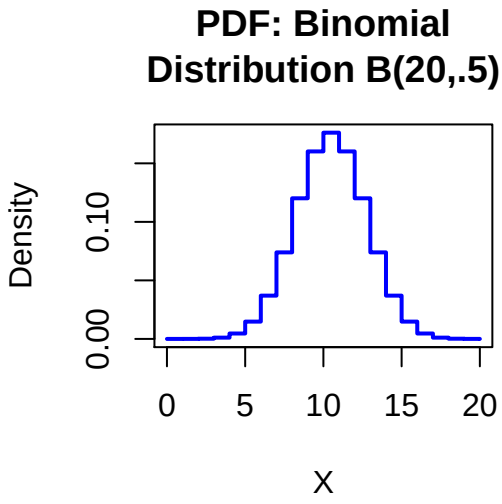
The PMF for the binomial distribution can be calculated as

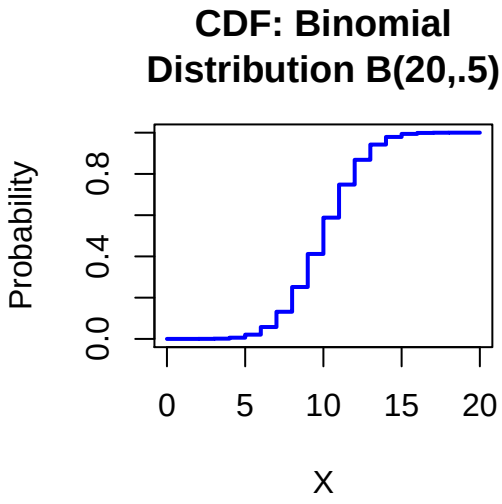
$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$.

$CDF = Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (n-i)^{n-i}$ where $\lfloor k \rfloor$ is the largest integer $\leq k$.

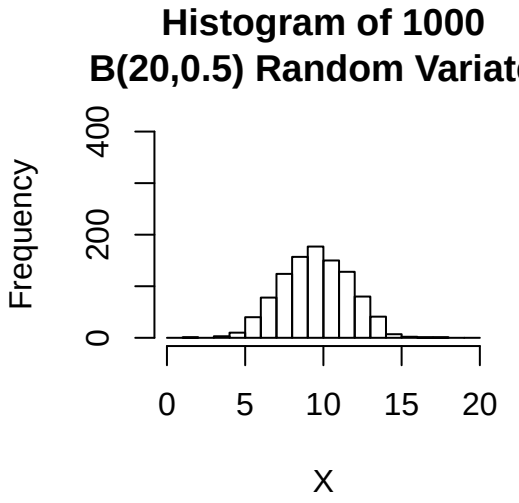
$Mean = np$

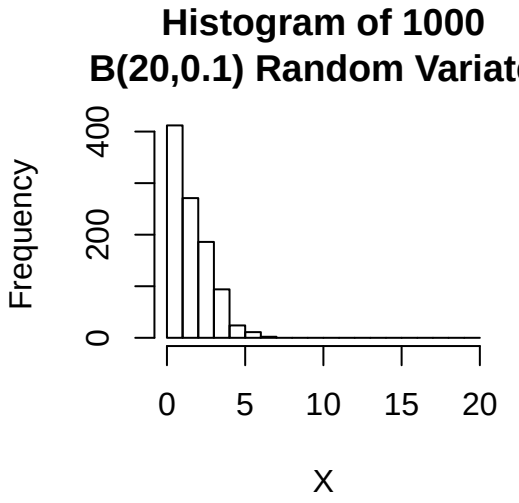
$Variance = np(1-p)$.





In Stata, binomial random variates can be generated using `generate x = rbinom(n,p)`, where n is the number of trials and p is the probability of success on each trial. The PMF for a specific outcome can be obtained using `binomial(n,k,p)`, where n is the number of trials (1 for Bernoulli), k is the number of successes, and p is the probability of a success on one trial. The CDF can be obtained using `binomial(n,k,p)`. $1 - CDF$ can be obtained using `binomialtail(n,k,p)`. Inverse functions are available in Stata as `invbinomial(n,k,p)` and `invbinomialtail(n,k,p)`.
Both the number of trials (n) and the probability of success (p) can affect the shape of the distribution, becoming more symmetric as the number of trials increases and as the success probability approaches 0.5.
Consider the expected distribution of heads obtained from flipping 20 fair coins ($p = .5$) and 20 biased coins ($p = .1$).





4.4 Poisson

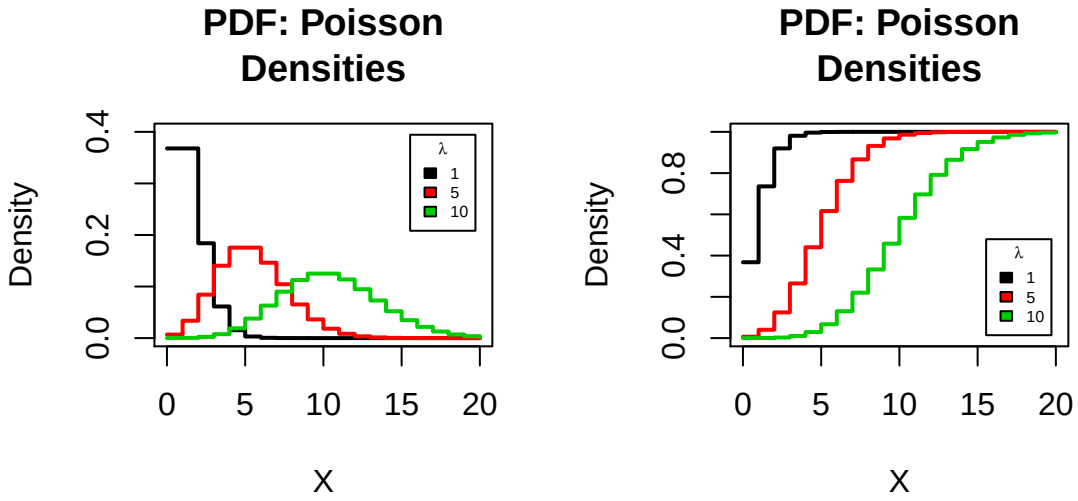
The Poisson distribution represents the number of events, k, that are expected within a given period of time for a given rate of occurrence, λ.

The PMF of a Poisson distribution is given by $PMF = \frac{\lambda^k}{k!}e^{-\lambda}$.

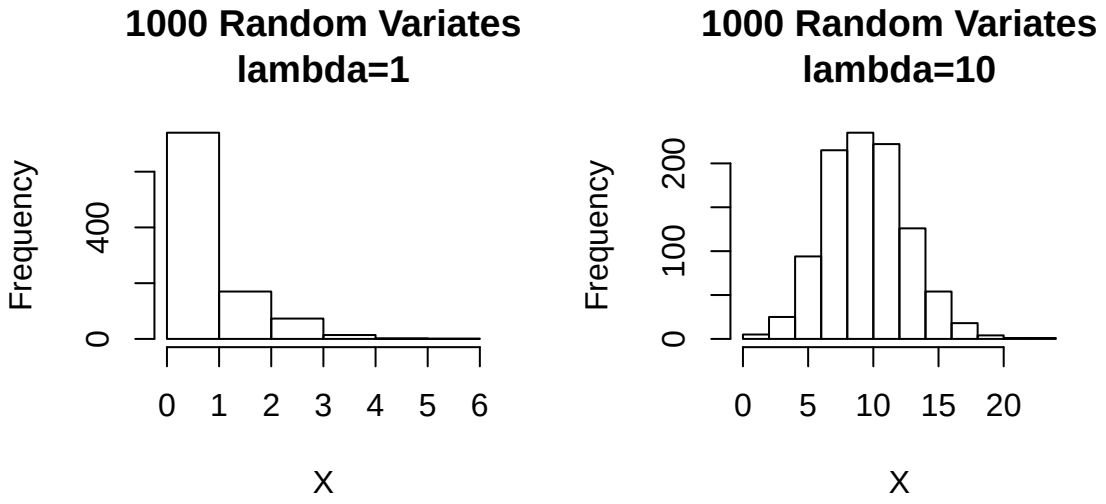
$CDF = e^{\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$

$Mean = \lambda$

$Variance = \lambda$



In Stata, Poisson random variates can be generated using `generate x = rpoisson(lambda)`, where lambda $= \lambda$. The PMF for a specific number of outcomes can be obtained via `poissonp(lambda,k)`, where lambda is λ and k is the number of events. The CDF can be obtained using `poisson(lambda,k)` and $1 - CDF$ can be obtained using `poissontail(m,k)`. Inverse functions are available as `invpoissontail(k,q)` and `invpoissontail(k,q)`, where q is the desired quantile.
Below are 1000 Poisson random variates drawn from populations with λ equal to 1 and 10, respectively.



4.5 Normal (Gaussian)

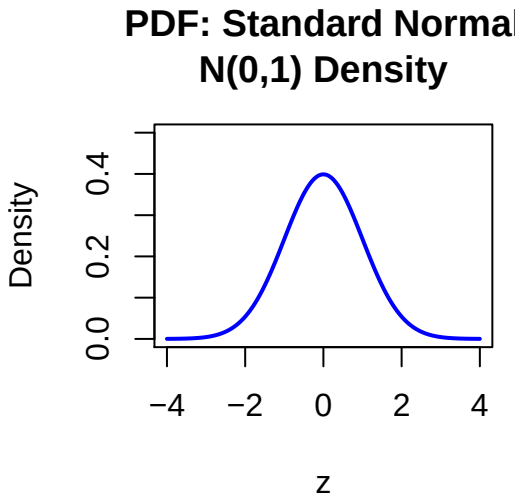
The normal, or Gaussian, distribution is nearly ubiquitous within statistical inference. For reasons that we will consider in this course, it shows up often, even when we may not expect it. The normal distribution is characterized by two parameters, the mean μ and variance σ^2, and is denoted by $N(\mu, \sigma^2)$. Standard normal variates are often referred to as z-scores (lowercase z). Observations drawn from $N(50, 100)$ are common in some applications such as education and personality assessment and are referred to as T-scores (uppercase T).

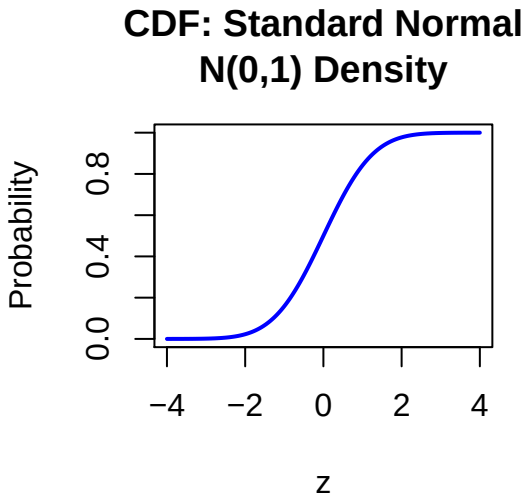
The CDF of a standard normal variate, $N(0, 1)$, is given as $PDF = \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$CDF = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$

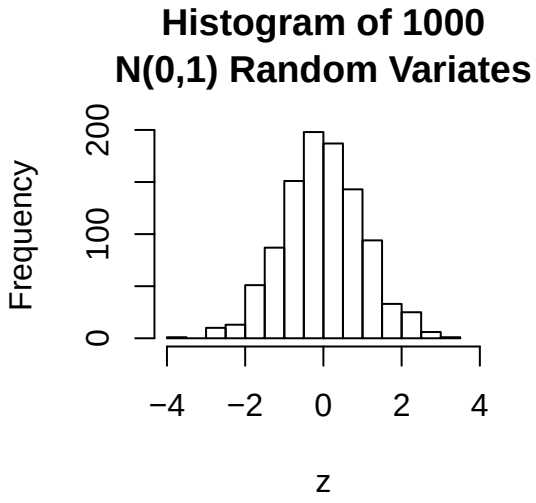
$Mean = \mu$

$Variance = \sigma^2$.





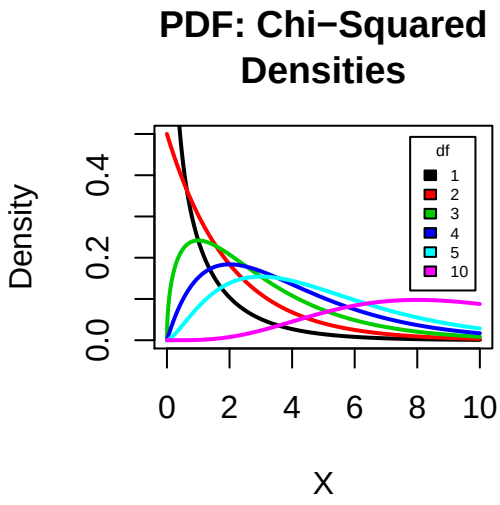
In Stata, normal random variates can be generated by using `generate x = rnormal(m,s)` where m is the mean and s is the standard deviation (i.e., the square root of the variance). The PDF for a normal distribution can be obtained via `normalden(x,m,s)`, where x is the value of interest, m is the distribution mean and s is the distribution standard deviation. The CDF is available through `normal(z)`, where z is a z-score. Note that for non-standard normal distributions, you will need to standardize your values prior to evaluation.
Likewise, inverse normal values are available only for the standard normal distribution via `invnormal(p)`, where p is the quantile of interest. Results are returned in standard normal metric.

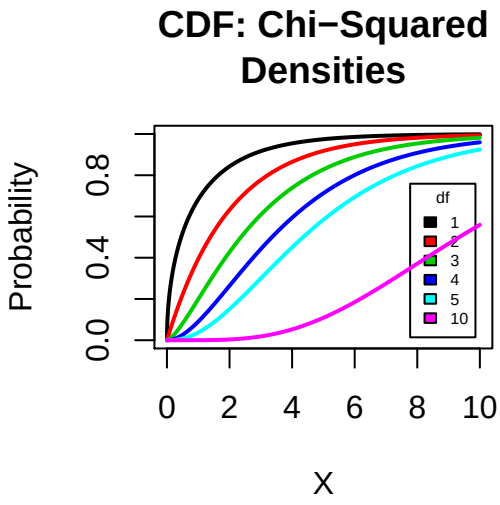


4.6 Chi-squared

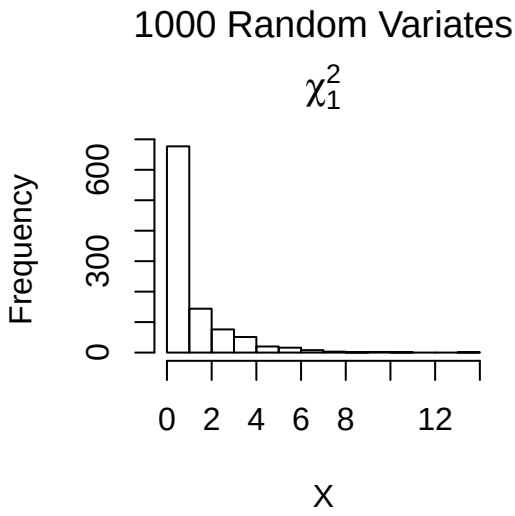
A chi-squared distribution is parameterized by its degrees of freedom, df, and is denoted by χ^2_{df}. It can be generated by squaring and summing df independent standard normal variates.

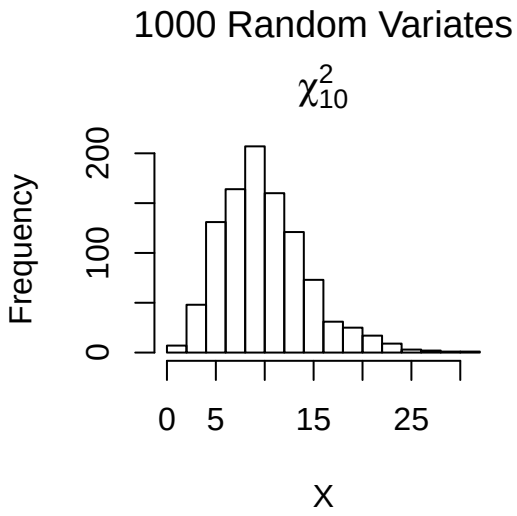
The PDF and CDF of a chi-squared distribution are a function of something called a Gamma (Γ) function, which is beyond the scope of our discussions here.
$Mean = df$
$Variance = 2df$.





In Stata, chi-squared random variates can be generated by using `generate x = rchi2(df)`, where df is the degrees of freedom. To estimate the PDF for a chi-squared distribution, use `chi2den(df,x)`, where x is the value of interest and df is df. To estimate the CDF, use `chi2(df,x)`. $1 - CDF$ can be estimated via `chi2tail(df,x)`. Chi-squared values for specific p-values can be obtained using `invchi2(df,p)` and `invchi2tail(df,p)`.

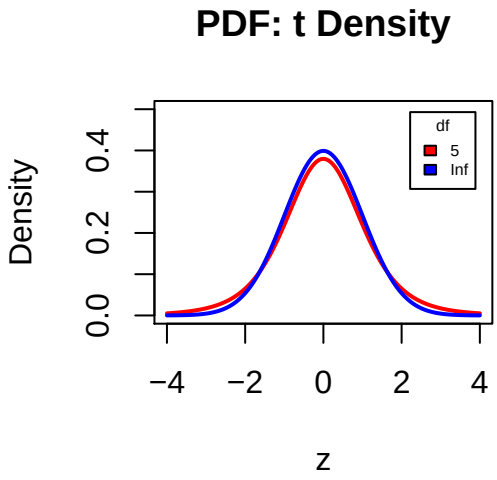


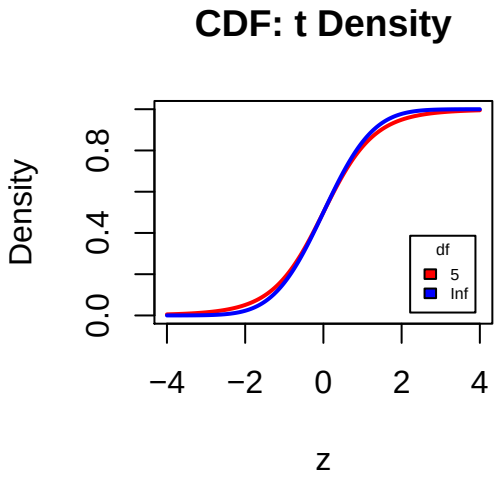


4.7 Student's t

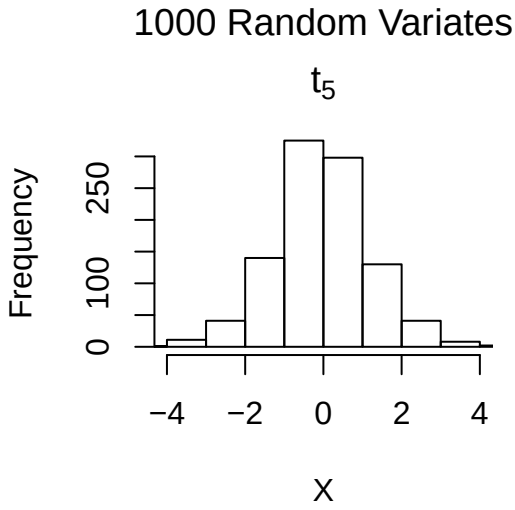
In contrast to normal distributions, which describe an entire population, t distributions describe samples, with small sample sizes where the standard deviations is unknown. The t

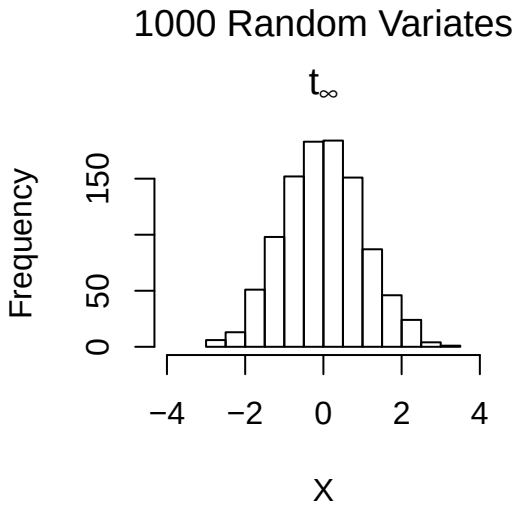
distribution is parameterized by its degrees of freedom, denoted as t_{df}. Compared with the normal distribution, the t distribution has more of the distribution in its tails. Like the χ^2 distribution, the PDF and CDF of the t distribution are a function of the Γ function.
$Mean = 0$
$Variance = \frac{df}{df-2}$, for $df > 2$.





In Stata, t random variates can be generated by using `generate x = rt(df)`, where df is the degrees of freedom. PDF for specific t-values can be obtained using `tden(df,t)`. CDF and $1 - CDF$ can be obtained via `t(df,t)` and `ttail(df,t)`, respectively. The t scores associated with specific p-values (and df) can be obtained via `invt(df,p)` and `invttail(df,p)`.





4.8 F

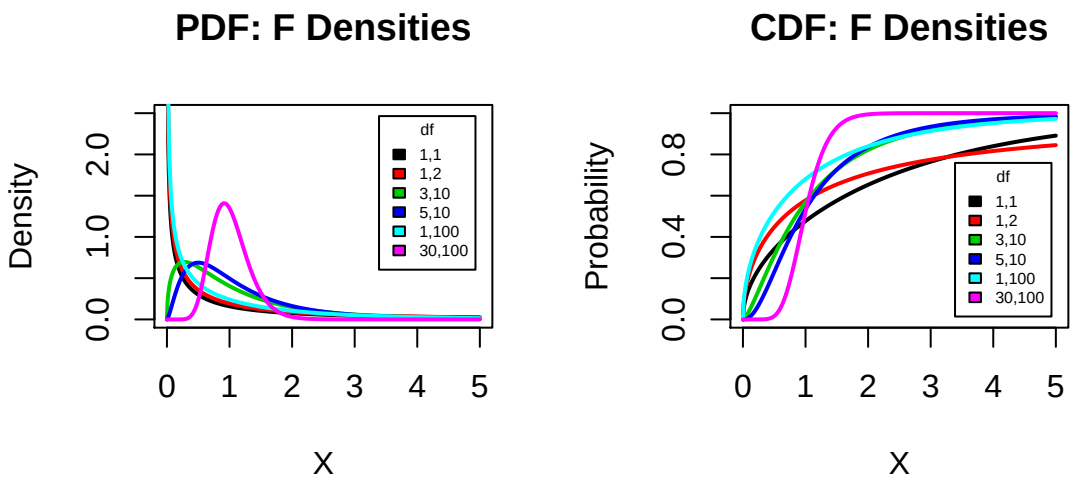
Many kinds of statistical models, such as ANOVA and multiple linear regression generate an F-statistic to evaluate the omnibus association between the set of predictors and the response variable. The F distribution has two parameters representing the numerator and denominator degrees of freedom: $F(df_n, df_d)$. The F distribution can be generated in terms of the ratio of two χ^2 distributions as follows.

$$F = \frac{\frac{\chi^2_n}{df_n}}{\frac{\chi^2_{df_d}}{df_d}}$$

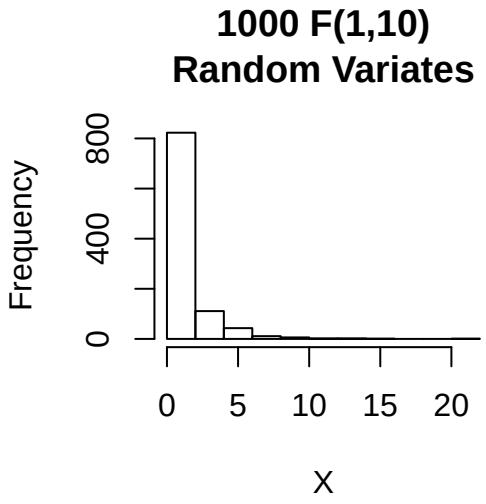
The PDF and CDF of the F-distribution are Beta functions, and are beyond the scope of this exercise.

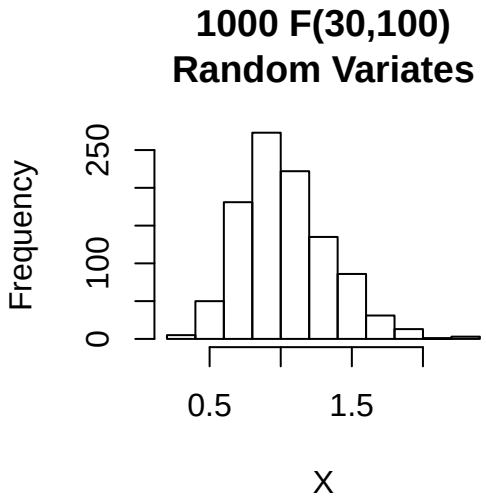
$$Mean = \frac{df_d}{df_d - 2}, \text{ for } df_d > 2$$

$$Variance = \frac{2df_d^2\left(df_n + df_d - 2\right)}{df_n\left(df_d - 2\right)^2\left(df_d - 4\right)}, \text{ for } df_d > 4$$



In Stata, F random variates can be generated by using `generate x = (rchi2(dfn)/dfn)/(rchi2(dfd)/dfd)`, where dfn is df_n and dfd is df_d.PDF for specific F-values can be obtained using `Fden(dfn,dfd,f)`. CDF and $1 - CDF$ can be obtained via `F(dfn,dfd,F)` and `Ftail(dfn,dfd,F)`, respectively. The F scores associated with specific p-values, df_n, and df_d can be obtained via `invF(dfn,dfd,p)` and `invFtail(dfn,dfd,p)`.


1000 F(1,10)
Random Variates
Frequency
0
400
800
0
5
10
15
20
X


1000 F(30,100)
Random Variates
Frequency
0
100
250
0.5
1.5
X

Stata Syntax: Distributions

```
/*****************************************************************
* Sample Distributions Syntax
* PH 8012, Spring 2015
* Adam Davey
*****************************************************************/

#delimit;
clear all;
capture log close;
log using "mylog.log", replace;

* Below simulates some data for the workflow;
set seed 12345;
set obs 1000;

* Uniform;
gen uniform = runiform();

* Bernoulli;
gen cointoss = rbinomial(1,0.5);
di binomialp(1,1,0.5);
di binomial(1,1,0.5);
di invbinomial(1,0,0.5);

* Binomial;
gen cointosses = rbinomial(20,0.5);
di binomialp(20,10,0.5);
di binomial(20,10,0.5);
di invbinomial(20,10,0.5);
di invbinomialtail(20,10,0.5);

* Poisson;
gen helminths = rpoisson(5);
di poissonp(5,4);
di poissontail(5,5);
di poisson(5,5);


* Normal;
gen iq = rnormal(100,16);
di normalden(0);
di normalden(100,100,16);
di normal(0);
```

```
di invnormal(0.5);
di invnormal(0.975);

* Chi-squared;
gen chi10 = rchi2(10);
di chi2den(10,10);
di chi2(1,3.84);
di chi2tail(1,3.84);
di invchi2(1,0.95);
di invchi2tail(1,0.05);

* t;
gen t5 = rt(5);
di t(5,0);
di ttail(5,0);
di tden(5,0);
di invt(5,0.975);
di invttail(5,0.025);

* F;
gen f30_100 = (rchi2(30)/30)/(rchi2(100)/100);
di F(30,100,1);
di Ftail(30,100,1);
di Fden(30,100,1);
di invF(30,100,0.95);
di invFtail(30,100,0.05);

summ, detail;
log close;

* Convert text output to PDF;
translate mylog.log mylog.pdf, replace;
```