Preliminary Statistics

Lecture 3: Probability Models and Distributions

Sophie van Huellen (sv8@soas.ac.uk), September 2013

Outline

- Probability Density Functions
- Cummulative Distribution Functions
- Expected Values
  - Properties
  - Variance
  - Covariance
- Common Continuous Models
  - Normal Distribution
  - T-Distribution
  - Chi-Square Distribution
  - F-Distribution
  - Sampling Distributions

Probability Densitiy Function

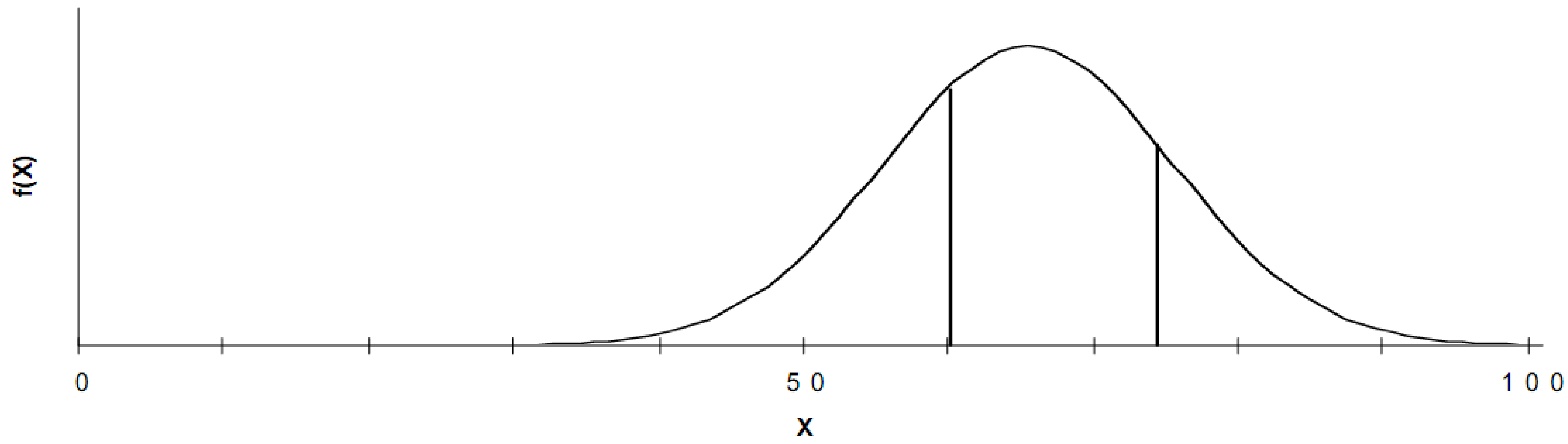
- A formula defining the curve of a continuous probability model.
- The area under the curve between two points gives the probability that a value between the two points will arise.
- This area is obtained by integrating the PDF between the two points.
- The sum of the area under the curve is equal to one.

Probability Densitiy Function (cont.)

- PDF of a continuous random variable measures the probability of the random variable over a certain range or interval.

$$f(x) = P(a \leq X \leq b) = \int_b^a f(X)dX$$

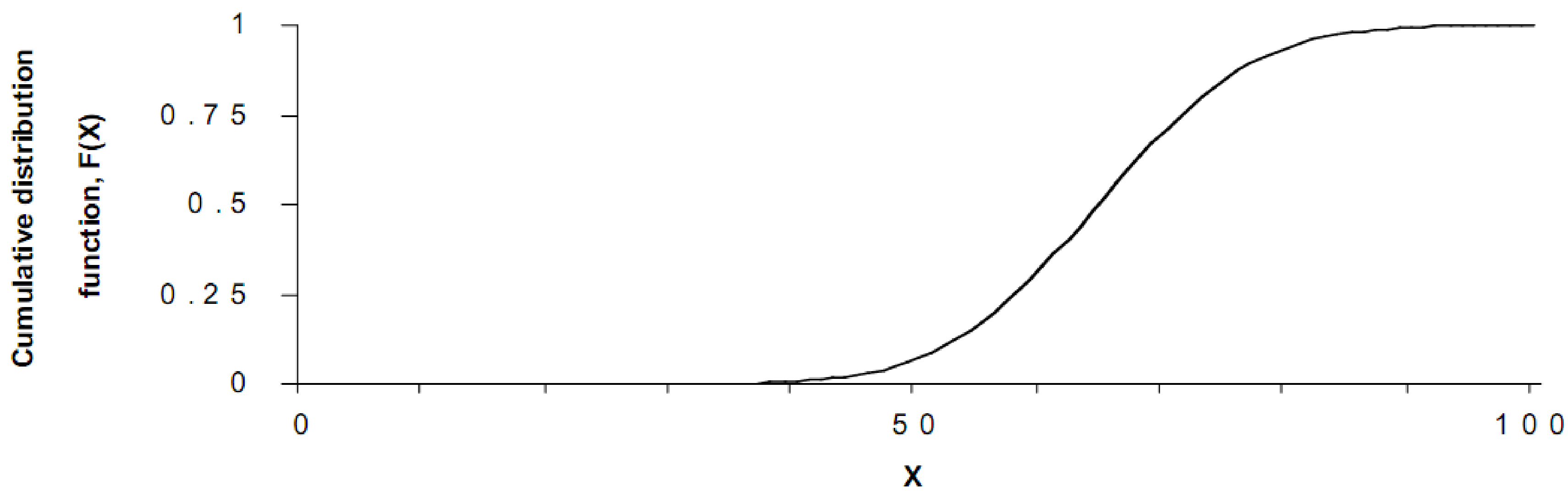
- For example, the probability that the height of an individual lies in the interval 60 and 75 inches is given by the area between 60 and 75.



Cummulative Distribution Function

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(X)dX$$

- Geometrically, the CDF of a continuous random variable would be a continuous curve



Characteristics of Distributions

Characteristics of Distributions			
Moments			
rth moment	$\frac{\sum_{i=1}^{N} x_i^r}{N}$	r=1 (the mean)	$\frac{\sum_{i=1}^{N} x_i}{N}$
Centered momentes			
rth moment	$\frac{\sum_{i=1}^{N} (x_i - \mu)^r}{N}$	r=2 (variance)	$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
Standardised moments			
rth moment	$\frac{\sum_{i=1}^{N} z_i^r}{N}$	r=3 (skewness) and r=4 (kurtosis)	

Expected Values

- For **discrete** random variables: $E(X) = \sum_{i=1}^{N} x_i f(x_i) = \mu$
- For **continuous** random variables: $E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \mu$
- The expected value of random variables is also known as the population mean. *E.g. In the example with the dice:*

$$E(X) = 7 = 2 * \frac{1}{36} + 3 * \frac{2}{36} + 4 * \frac{3}{36} \ldots + 12 * \frac{1}{36}$$

- If all values are equalliy likely, $f(x_i)=1/N$, and the ecpected balue is the arithmetic mean.

Expected Values (cont.)
Properties

Properties of Expected Value		
1.	$E(b) = b$	Where b is a constant.
2.	$E(X+Y) = E(X) + E(Y)$	Where X and Y are random variables.
3.	$E(XY) \neq E(X)\,E(Y)$	Where X and Y are non independent random variables.
4.	$E(XY) = E(X)\,E(Y)$	If X and Y are independent random variables.
5.	$E(aX) = aE(X)$	Where a is a constant.
6.	$E(aX+b) = aE(X) + E(b) = aE(X) + b$	Where a and b are constant.

Expected Values (cont.)

Variance

- Let X be a **discrete** random variables with $E(X) = \mu$:

$$Var(X) = \sigma^2 = E(X-\mu)^2 = \sum_{i=1}^{N}(x_i-\mu)^2 f(x_i)$$

- The variance of X is the expected value of the square difference between values of X and its expected value.

- Let X be a **continuous** random variable with $E(X) = \mu$:

$$Var(X) = \sigma^2 = E(X-\mu)^2 = \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx$$

- If $f(x)=1/N$, this is the same as the population variance.

Expected Values (cont.)
Variance

	Properties of Variance	
1.	$Var(x) = E(X-\mu)^2 = E(X^2) - \mu^2$	
2.	$Var(b) = 0$	With b being a constant.
3.	$Var(X+b) = Var(X)$	With b being a constant.
4.	$Var(aX) = a^2\,Var(X)$	With a being a constant
5.	$Var(aX+b) = a^2\,Var(X)$	With a and b being constants
6.	$Var(X+Y) = Var(X) + Var(Y)$ $Var(X-Y) = Var(X) + Var(Y)$ $Var(aX+bY) = a^2\,Var(X) + b^2\,Var(Y)$	**If** X and Y are independent random variables. Else see properties of covariance. With a and b being constants.

Expected Values (cont.)

Covariance

- Let X and Y be two random variables with means $E(X) = \mu_x$ and $E(Y) = \mu_{y.}$
- Then the covariance between the two is:

$$cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

$$= \sum_X \sum_Y (X-\mu_x)(Y-\mu_y) f(X,Y) = E(XY) - \mu_x \mu_y$$

$$= \sum_X \sum_Y XYf((X,Y) - \mu_x \mu_y$$

Expected Values (cont.)
Covariance

Properties of Covariance		
1.	$cov(X,Y) = E(XY) - E(X)E(Y)$	If X and Y are two random variables
2.	$cov(X,X) = var(X)$	
3.	$cov(X,a) = 0$	With a being a constant.
4.	$cov(a+bX,c+dY) = bd\ cov(X,Y)$	With a,b,c, and d being constants
5.	$cov(X,Y) = 0$	If X and Y are independent, since $E(XY)=E(X)E(Y)=\mu_x\mu_y$
6.	$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$ $Var(X-Y) = car(X) + var(Y) - 2cov(X,Y)$	If X and Y are two random variables

Common Continuous Models
Normal Distribution

- Normality arises:
  - When a random variable is the result of many independent, random influences, none of which is dominant,
  - From the *Central Limit Theorem*,
  - When a random variable is loged.

Common Continuous Models (cont.)

Normal Distribution

- Probability Density Function of a Normal Distribution

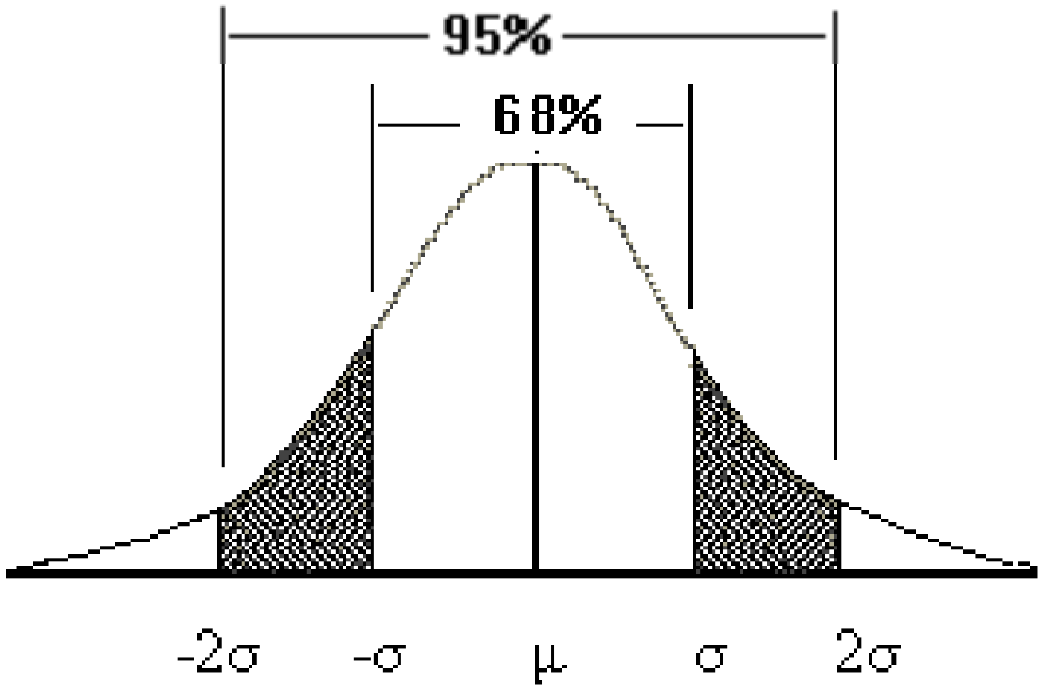
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- Two parameters: mean (μ) and variance (σ)
- The normal distribution is the exponential of a quadratic. $(2\pi\sigma^2)^{-\frac{1}{2}}$ makes it integrate (add up) to 1.
- $Y \sim N(\mu,\sigma^2)$: The random variable X is normally distributed with mean μ and variance σ^2.

Common Continuous Models (cont.)

Normal Distribution

- Properties:
  - Bell shaped and symmetric.
  - Mean = Median = Mode.
  - Skewness and Excess Kurtosis are equal to zero.



Common Continuous Models (cont.)
Normal Distribution

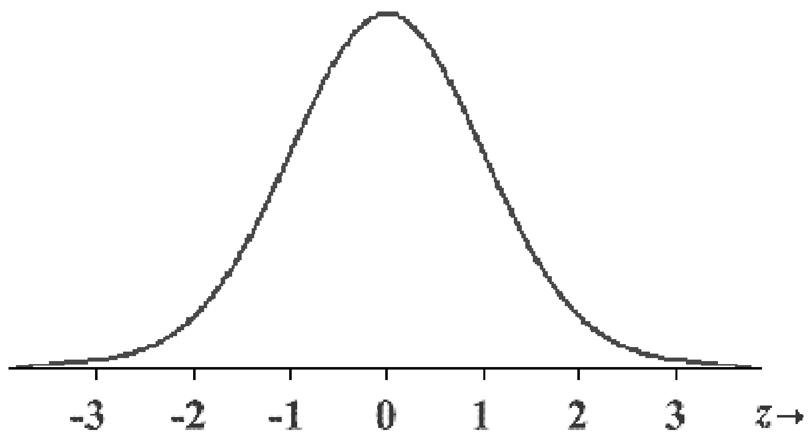
- Linear Transformations
  - Any linear transformation of a normally distributed random variable would also be normally distributed.
    - $Y \sim N(\mu, \sigma^2)$
    - $W = a + bY \sim N(a + b\mu, b^2\sigma^2)$
  - A very useful linear transformation is:

$$z_i = \frac{y_i - \mu}{\sigma}$$

Common Continuous Models (cont.)

Standard Normal Distribution

- The standard normal distribution has mean zero and variance (s.d) one.
- $Z \sim N(0,1)$
- The SND is the reference distribution for the normal tables.



Common Continuous Models (cont.)

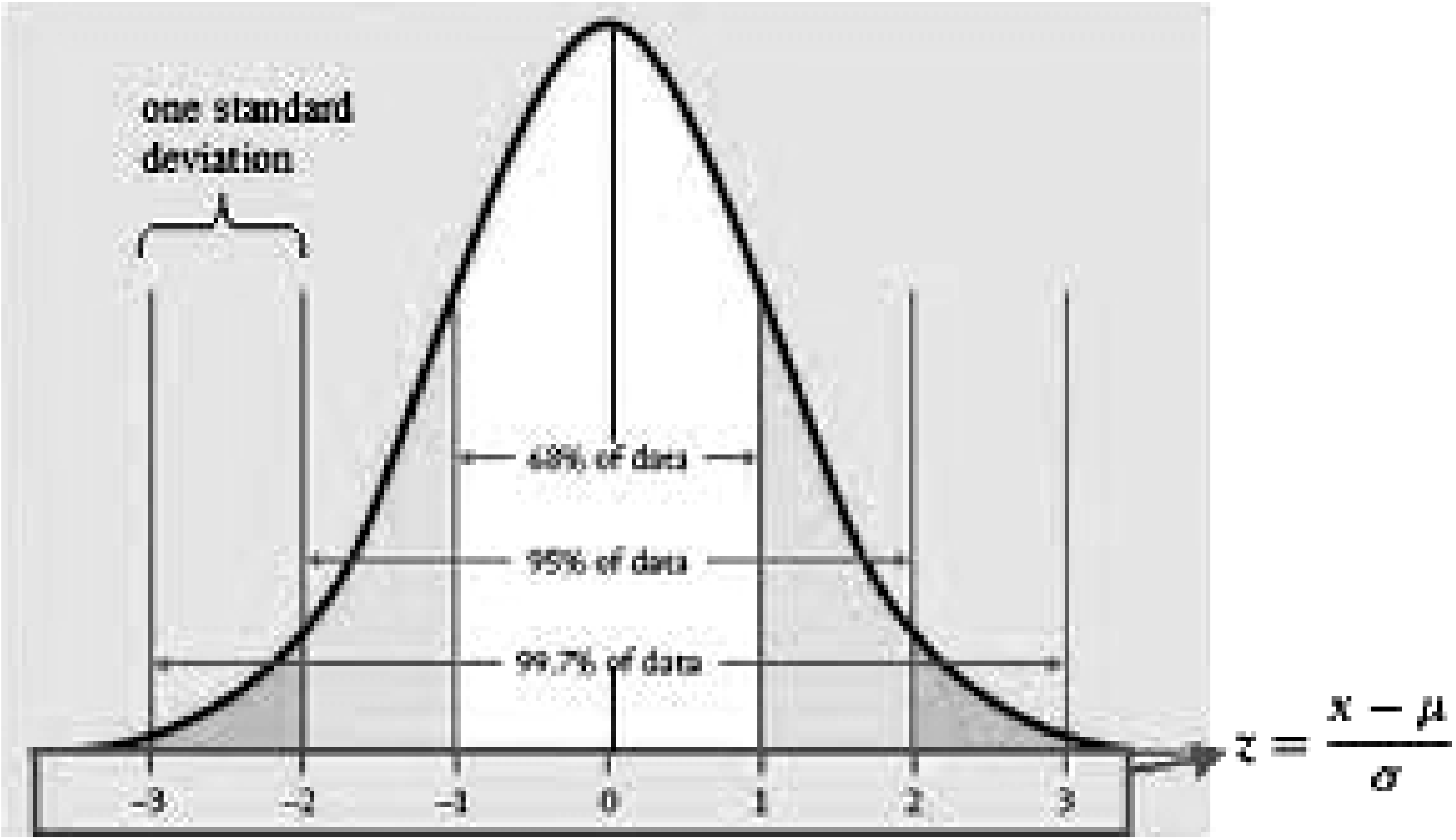
Standard Normal Distribution

Areas Under a Normal and Standard Normal Distribution

Z	$Y = \mu + \sigma Z$	Probabilities	What it means
$P(-1 < Z < 1)$	$P(\mu - \sigma < Y < \mu + \sigma)$	= 0.6826	Prob. of being within 1 sd From mean is 68.26%.
$P(-1.96 < Z < 1.96)$	$P(\mu - 1.96\sigma < Y < \mu + 1.96\sigma)$	= 0.96	95% of the normal distribution lies within 1.96 sds form the mean.
$P(-2 < Z < 2)$	$P(\mu - 2\sigma < Y < \mu + 2\sigma)$	= 0.9544	Prob. of being within 2 sd from the mean is 95.44%.
$P(-3 < Z < 3)$	$P(\mu - 3\sigma < Y < \mu + 3\sigma)$	= 0.997	Prob. of being within 3 sd from the mean is 99.7%.

Common Continuous Models (cont.)

Standard Normal Distribution



Common Continuous Models (cont.)

Distributions Related to Normal: Chi-Squared Distribution

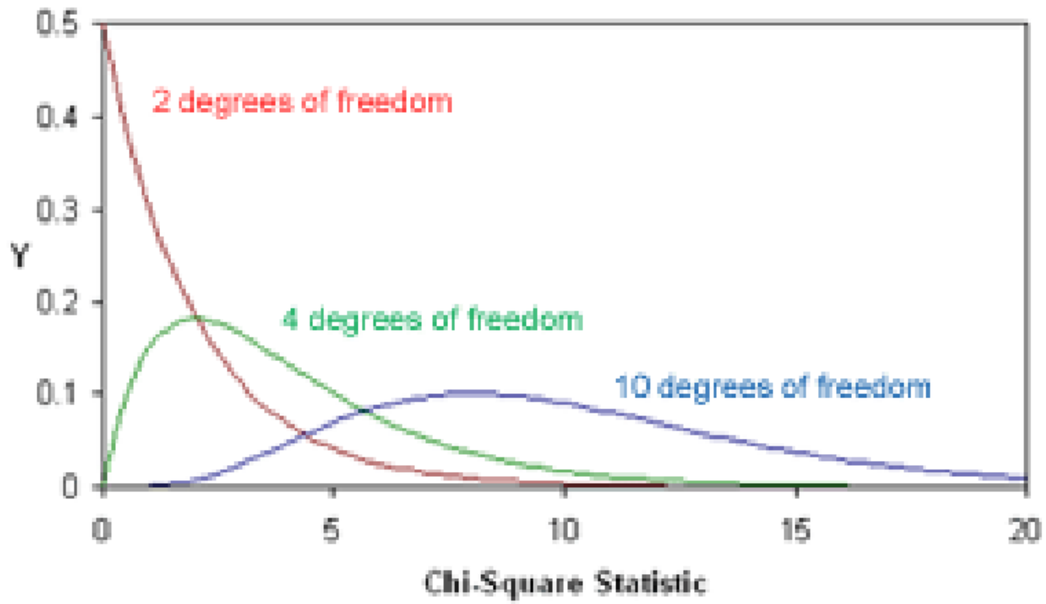
$$A = \sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$$

- The sum of n independent squared standard normal variables, $\sum z_i^2$ follows a chi-squared distribution with n degrees of freedom
- The chi-squared distribution takes only positive values and ranges from 0 to infinity

Common Continuous Models (cont.)

Distributions Related to Normal: Chi-Squared Distribution

- χ^2 is skewed.
- For relatively few degrees of freedom (d.f.) the distribution is highly skewed to the right, but as the d.f. increase, the distribution approaches the normal.
- The mean of the chi-squared random variable is n and its variance is 2n, where n is the d.f.



Common Continuous Models (cont.)

Distributions Related to Normal: Student‘s t-Distribution

$$t(n) = z / \sqrt{\frac{\chi^2(n)}{n}}$$

- The Student‘s t-Distribution is a bell shaped, symmetric distribution, similar to the standard normal distribution.
- Used to compensate for the extra uncertainty when the sample is used to calculate the parameters of a normal distribution.

Common Continuous Models (cont.)

Distributions Related to Normal: F-Distribution

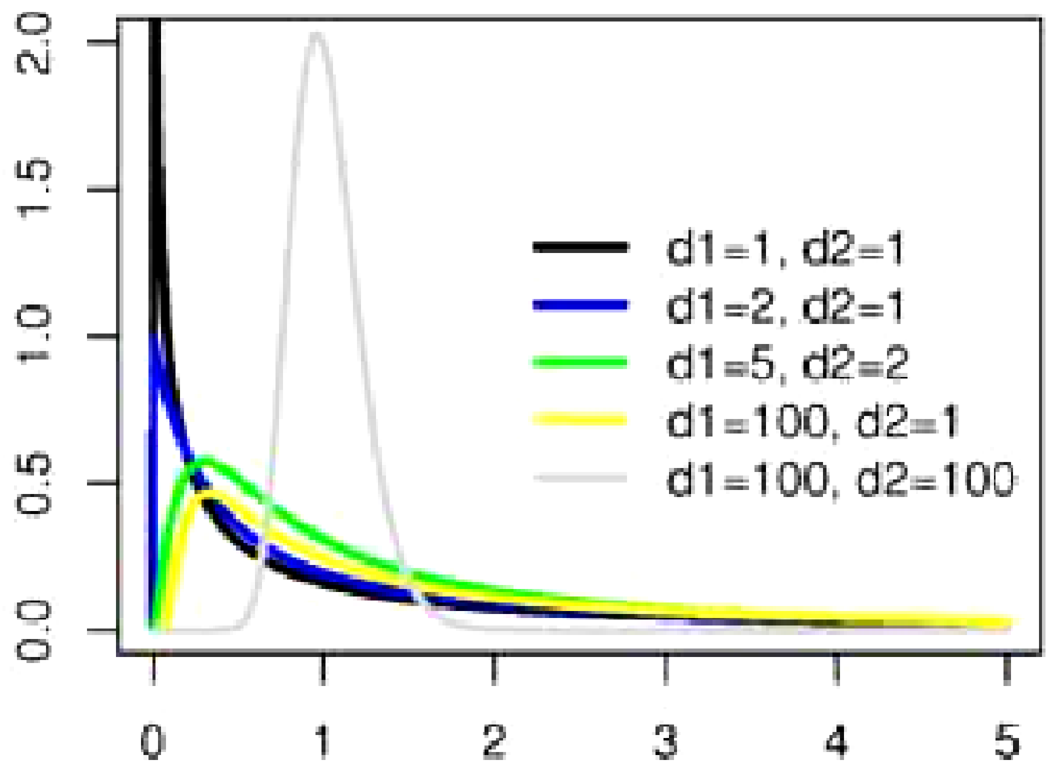
$$F_{n,m} = \frac{\chi_n^2 / n}{\chi_m^2 / m}$$

- The ratio of two chi-squared variables, each divided by its degrees of freedom follows an F-distribution.
- Two indexing parameters: the degrees of freedom in the numerator and the degrees of freedom in the denominator.

Common Continuous Models (cont.)

Distributions Related to Normal: F-Distribution

- F resembles the χ^2 : is always non negative and is skewed to the right.



Common Continuous Models (cont.)

Distributions Related to Normal: F-Distribution

Left-Hand Tail of an F-Distribution	
Use the identity	$F^{c}_{b,1-\alpha} = \dfrac{1}{F^{b}_{c,\alpha}}$
E.g.	
For $F_{5,10}$ the 0.05 right-hand critical value	$F^{5}_{10,0.05} = 3.33$
For $F_{10,5}$ the 0.05 right-hand critical value	$F^{10}_{5,0.05} = 4.74$
For $F_{5,10}$ the 0.05 left-hand critical value	$F^{5}_{10,0.95} = \dfrac{1}{4.74} = 0.21097$

Common Continuous Models (cont.)
Sampling Distributions

- Given random sampling.
- Statistics, being calculated from a sample, are random variables.
- The value of the statistic is the outcome of a random process, the random sampling process.
- The value of the statistic will differ depending on the sample drawn.
- As such it will have a spread of values, each with an associated probability, i.e. a probability distribution.
- Probability distributions for statistics are commonly referred to as sampling distributions.