Quantitative Analysis of Financial Markets

Session 3 Pre-class preparations: Statistical Distributions

Benjamin Ee

SMU

SINGAPORE MANAGEMENT UNIVERSITY

Masters of Quantitative Finance

QF 603

Week 3

Pre-class Prep 1/11

Video Introduction, not more than 30 seconds

Introduction and Instructions

1. 11 videos for this week's set of mandatory pre-class preparation
2. Each video is around 6 to 10 minutes long
3. You can watch the videos as few or as many times as you like
4. For some videos, after you finish watching them, it may direct you towards quizzes in eLearn that you have to do.
5. For each quiz, note that the system is programmed to allow 4 attempts
6. Complete all quizzes before the start of next class. Performance will automatically flow through to HW component of this course
7. Expected takeaways: You are not expected to memorize derivations or formulas except for selected results.
8. Look out for this icon next to material / formulas that you should try to know well for the rest of this term

Why do we need to study statistical distributions?

1. **Confidence Intervals, Interference and Hypothesis testing**
    a) In class 2, we studied confidence intervals using a normal distribution. Confidence intervals are important for forecasting
    b) Not everything follows a normal distribution. We need to also familiar with other distributions

2. **Risk management**
    a) Risk management is concerned with the distribution of tail events. This allows us to compute value at risk (VAR). To know the probability of tail events, we need to know the shape of the distribution, which does not have to be normal

See video

Recap on “what are statistical distributions”

1. With any dataset, the data values will follow a specific frequency distribution
2. The frequency distribution may also follow a certain ‘shape’, which can tell us “what are the chances of any random observation in the dataset falling within a certain range”
3. Therefore, one formal definition of a statistical distribution is that it should have a probability density function (for continuous) or a probability mass function (for discrete)

See video

Mean and Variance of (any) Distributions

Week 3
Pre-class Prep 2/11

Mean and Variance of statistical distributions

- As we study various distributions, one goal will be to understand how the **mean and variance** of the distribution evolves with the “parameters” of the distribution.

- We focus on mean and variance because they are often used for hypothesis testing in finance. For instance, one hypothesis test could be that the mean is large compared to the variance

- Because we need these quantities, we will want to express mean and variance in terms of the parameters of the distribution

- For example, a binomial distribution involves a “coin flip”. It has a parameter p, which denotes the probability of “heads” (and 1-p denotes probability of “tails”). We can then ask “**how does changing p affect the mean and variance of the binomial distribution**?”

Mean of a statistical distribution

- Let the probability density function (pdf) of the distribution be f(x)
- For a discrete variable, recall that the pdf of a distribution f(x) tells us the point probability of x.
- For a continuous variable, the pdf of the distribution allows us to derive the probability that x will fall in between two limits
- The mean of any distribution is therefore given by: $E[X] = \int_{-\infty}^{\infty} x f(x) dx$
- Note that for a discrete distribution, we can replace the integration sign with a summation sign

Example: computing mean of distribution

See video

Variance of a statistical distribution

- Denoting f(x) as the pdf of a distribution (same as previous slide), we can compute variance of any distribution as:

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx,$$

- Similar to before, we can replace the integration sign with a summation sign for a discrete distribution

Example computation for variance of distribution

$f(x) = \frac{8}{9}x \; for \; 0 \leq x \leq 1.5$

Variance of x:

1. $\int_0^{1.5}(x-u)^2 f(x)dx = \int_0^{1.5}(x-1)^2 f(x)dx \;\; where \; u = 1 \; was \; obtained \; previously$
2. $\int_0^{1.5}(x-1)^2 f(x)dx = \int_0^{1.5}(x-1)^2 \frac{8}{9}xdx$
3. $\int_0^{1.5}(x-1)^2 \frac{8}{9}xdx = \int_0^{1.5}\frac{8}{9}(x^3 + x - 2x^2)dx$
4. $\int_0^{1.5}\frac{8}{9}(x^3 + x - 2x^2)dx = \frac{8}{9}[\frac{x^4}{4} + \frac{x^2}{2} - 2\frac{x^3}{3}]$ evaluated at 1.5 and 0; Let F(x) be the integral
5. F(1.5) – F(0) = 1/8

Bernoulli Distribution

Week 3
Pre-class Prep 3/11

3. BERNOULLI

- A Bernoulli random variable X is either zero or one.
- If we define p as the probability that X equals one, we have

$$\mathbb{P}(X=1)=p; \quad \mathbb{P}(X=0)=1-p.$$

- A Bernoulli r.v. is analogous to a **single** flip of a coin. Outcome is determined by probability the coin is heads or tails

3. BERNOULLI

- Binary outcomes are common in finance: a bond can default or not default; the return of a stock can be positive or negative; a central bank can decide to raise rates or not to raise rates.

3. BERNOULLI

- The mean and variance of a Bernoulli random variable can be computed using same formulas as before

$$\mu = p \times 1 + (1-p) \times 0 = p$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \sum_{x_i=0}^{1} x_i P(x = xi) = (1 \times P(x=1)) + (0 \times P(x=0)) = 1 \times p + 0 \times (1-p) = p$$

$$\sigma^2 = p \times (1-p)^2 + (1-p) \times (0-p)^2 = p(1-p).$$

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx,$$

$$= \sum_{x_i=0}^{1} (x_i - p)^2 P(x = xi)$$

$$= \left((1-p)^2 \times P(x=1)\right) + \left((0-p)^2 \times P(x=0)\right) = (1-p)^2 p + (0-p)^2 (1-p)$$

$$= p(1-p)$$

Bernoulli Distribution is discrete with p being the only parameter.

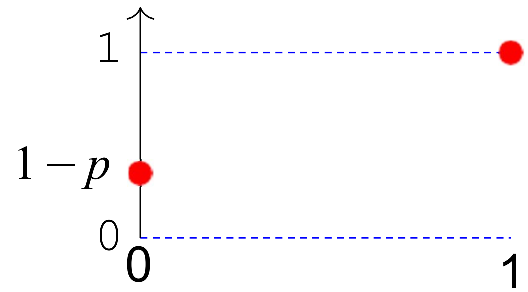
PDF and CDF of Bernoulli

- The probability of a Bernoulli random variable can be written as

$$P(X = x) = p^x(1 - p)^{1-x}.$$

- The cumulative distribution function is

$$F(X = x) = \begin{cases} 1 - p, & \text{for } x = 0; \\ 1, & \text{for } x = 1. \end{cases}$$



Estimation of p

- How can we estimate p of a coin for the head to turn up?
- Throw N times, and count the number (n) of times the head has turned up. Then

$$\widehat{p} = \frac{n}{N}.$$

- Is this a good estimator?

Binomial Distribution

Week 3
Pre-class Prep 4/11

4. BINOMIAL

- The binomial distribution gives the discrete probability distribution $P(n, N; p)$ of obtaining exactly n successes out of N Bernoulli trials.
- Binomial distribution pdf is given by: $$\mathbb{P}(n;N,p) = \binom{N}{n} p^n (1-p)^{N-n}.$$

See video

4. BINOMIAL

- The cumulative distribution function is

$$F(k;N,p) = \mathbb{P}(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{N}{i} p^i (1-p)^{N-i},$$

where *upper limit* is the largest integer <= k.

We can derive above formula via **observation that outcomes are mutually exclusive.**

Example: Either you have **6 heads out of 10 flips, or 7 heads**. You cannot have both in the same sequence.

From N “coin flips”, **probability that number of “heads” is <= C** (where C is an integer) is:

P(#H <= C) = P(H= 1) + P(H=2) + P(H=3) + …. + P(H = C)

This explains summation sign above

Estimating p

- If we do not know *p*, we can estimate from empirical data *p* = *n/N*
- To show that n/N is an 'unbiased estimator', we need to show that E(n/N) = p
- i.e. $\int_{n=0}^{N} \frac{n}{N} f(x_i = n)dx = \sum_{n}^{N} \frac{n}{N} \binom{N}{n} p^n (1-p)^{N-n}$

$$\mathbb{E}\left(\widehat{p}\right) = \sum_{n=0}^{N} \frac{n}{N} \binom{N}{n} p^n (1-p)^{N-n} = p \sum_{n=1}^{N} \frac{N!}{(N-n)!n!} \frac{n}{N} p^{n-1} (1-p)^{N-n}$$

$$= p \sum_{n=1}^{N} \frac{(N-1)!}{(N-n)!(n-1)!} p^{n-1} (1-p)^{(N-1)-(n-1)}$$

$$= p(1-p)^{N-1} \sum_{m=0}^{N-1} \binom{N-1}{m} \left(\frac{p}{1-p}\right)^m$$

$$= p(1-p)^{N-1} \left(1 + \frac{p}{1-p}\right)^{N-1} = p(1-p)^{N-1} \left(\frac{1}{1-p}\right)^{N-1} = p.$$

Unbiasedness proof for information, not examinable

Mean and Variance of Binomial distribution

1. To solve for **mean and variance of the binomial distribution**, we start in the same way as with Bernoulli distribution (or any other distribution) by using these formulas:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx,$$

2. Since Binomial is a discrete distribution, replace **integration signs above with summation signs**

3. Recall once again that we are primarily interested in “mean” and “variance” because both of these values play a **significant role in hypothesis testing**

Mean and Variance of Binomial distribution

1. Applying the 2 formulas mentioned on the previous slide,
2. Mean of Binomial distribution = **Np** where N is the number of trials
3. Variance of Binomial distribution = **Np(1-p)**

Poisson Distribution

Week 3
Pre-class Prep 5/11

5. POISSON DISTRIBUTION

1. For a Poisson random variable X, where λ is a constant parameter.

$$\mathbb{P}(X = n) = \frac{\lambda^n}{n!} e^{-\lambda},$$

2. The Poisson distribution is often used to model
   - the occurrence of events over time such as the number of bond defaults in a portfolio
   - the number of crashes in equity markets.
   - jumps in jump-diffusion models

Do not have to memorize the pdf formula, need to know how to apply

Poisson Distribution from Binomial Model

1. Recall that the probability of obtaining exactly n successes in N Bernoulli trials is given by

$$\mathbb{P}\left(X = n; N, p\right) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}.$$

2. Viewing the distribution as a function of the expected number of successes $\lambda := Np$ instead of the sample size N for fixed p, we write

$$\mathbb{P}\left(X = n; N, \lambda := Np\right) = \frac{N!}{n!(N-n)!} \left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n}.$$

<u>This is just a relabelling of the variable p!</u>

Poisson Distribution from Binomial Model

(cont'd)

1. What happens if the sample size N becomes large, even infinite?

$$
\begin{aligned}
P_\lambda(n) &:= \lim_{N\to\infty} \mathbb{P}\left(X = n; N, \lambda := Np,\right) \\
&= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{n!} \frac{\lambda^n}{N^n} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-n} \\
&= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{N^n} \frac{\lambda^n}{n!} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-n} \\
&= 1 \times \frac{\lambda^n}{n!} e^{-\lambda} \times 1 \\
&= \frac{\lambda^n}{n!} e^{-\lambda}
\end{aligned}
$$

For background, not examinable

2/ Poisson distribution is a limiting case of binomial model.

Mean and Variance of Poisson Random Variable

1. First we show that $P_\lambda(n)$, $n = 0, 1, 2, \ldots, \infty$ indeed adds up to 1.

$$\sum_{n=0}^{\infty} P_\lambda(n) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1.$$

2. Mean is (once again) based on $E[X] = \int_{-\infty}^{\infty} x f(x) dx$

$$\sum_{n=0}^{\infty} n P_\lambda(n) = e^{-\lambda} \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} = e^{-\lambda} \lambda \sum_{n=0}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} = \lambda.$$

3. Similarly, variance is based on $Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx,$

4. Skipping the derivation, variance of Poisson R.V is (also) λ

Poisson Process

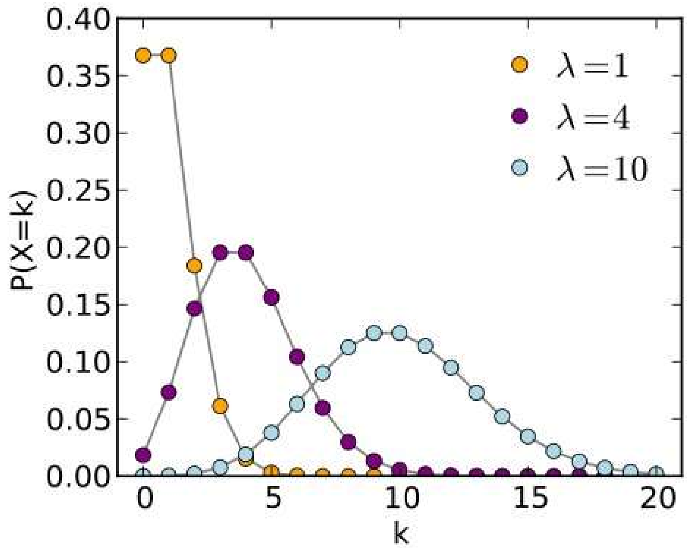
1. If the rate at which events occur over time is constant, and the probability of any one event occurring is independent of all other events, then we say that the events follow a Poisson process:

$$\mathbb{P}\left(X := N(t+\tau) - N(t) = n\right) = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau}.$$

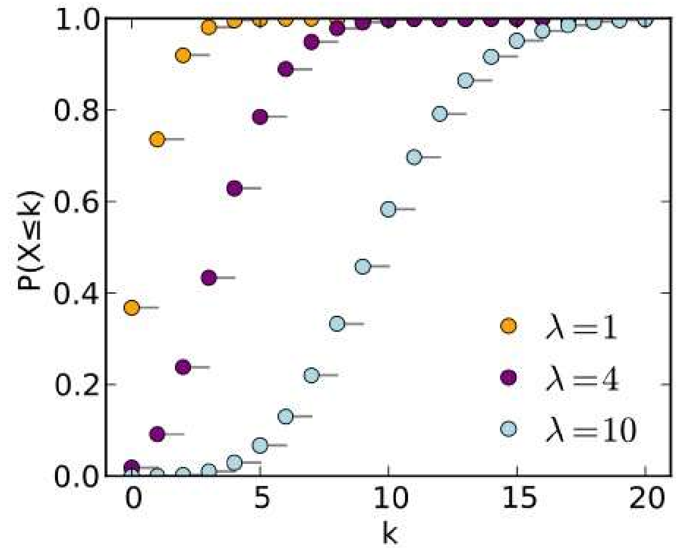
Here, $N(t+\tau) - N(t) = n$ is the number of events in the time interval $(t, t+\tau)$

2. The mean and variance of a Poisson process are the same: λt.

Probability Mass Function and CDF



Probability Mass Function



Cumulative Distribution Function

Source: https://en.wikipedia.org/

Sample Problem

1. Assume that defaults in a large bond portfolio follow a Poisson process. The expected number of defaults each month is four. What is the probability that there are exactly three defaults over the course of one month? Over two months?

2. Over one month, the probability is

$$\mathbb{P}\left(X = 3\right) = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau} = \frac{(4 \times 1)^3}{3!}e^{-4\times 1} = 19.5\%.$$

3. Over two months, the probability is

$$\mathbb{P}\left(X = 3\right) = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau} = \frac{(4 \times 2)^3}{3!}e^{-4\times 2} = 2.9\%.$$

Uniform Distribution

Week 3
Pre-class Prep 6/11

6. UNIFORM

1. So far, we have been talking about discrete distributions.

2. The continuous uniform distribution is first continuous distribution we are discussing, and one of the simplest

3. For all $x \in [b_1, b_2]$,

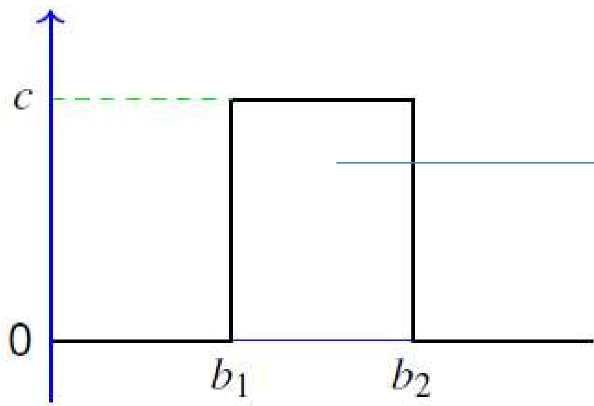
$$u(b_1, b_2) = c.$$

In other words, the probability density is constant and equal to c between b_1 and b_2, and zero everywhere else.

4. Because the probability of any outcome occurring must be one, you can find the value of c as____.

Clue: Total probability must be 1, so total area bounded by the pdf = 1

UNIFORM distribution: value of c



Area of this rectangle is 1

Therefore, c = 1/(b2 – b1)

Mean, Variance, PDF, and CDF

- The mean of the uniform random variable is

  <midpoint of b1 and b2>
  ______________________________.

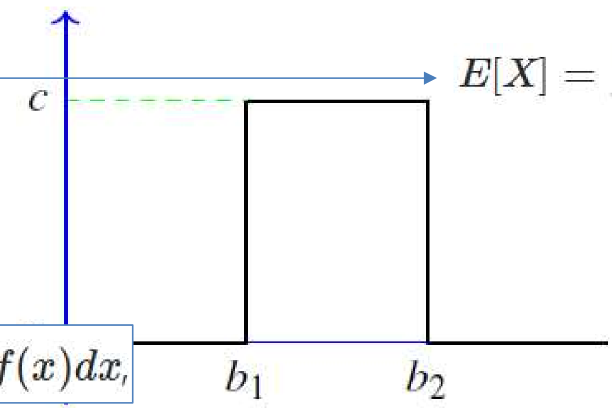
  $$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- The variance of the uniform random variable is

  $$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx,$$

  ____

- Probability Density Function



- The cumulative distribution function is

  $$F(a) = \int_{-\infty}^{a} f(x)\, dx = \mathbb{P}(X \leq a).$$

  $$\frac{(x - b1)}{(b2 - b1)}$$

Variance of Uniform Random Variable

$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx,$

1. $\int_{b1}^{b2} (x \; - \left(\frac{b1+}{2}\right))^2 f(x)dx =$
2. $\int_{b1}^{b2} (x \; - \left(\frac{b1+b2}{2}\right))^2 c \; dx =$
3. k
4. $\int_{b1}^{b2} cx^2 + c\left(\frac{b1+b2}{2}\right)^2 - cx(b1+b2)dx$ where c = 1/(b2-b1)

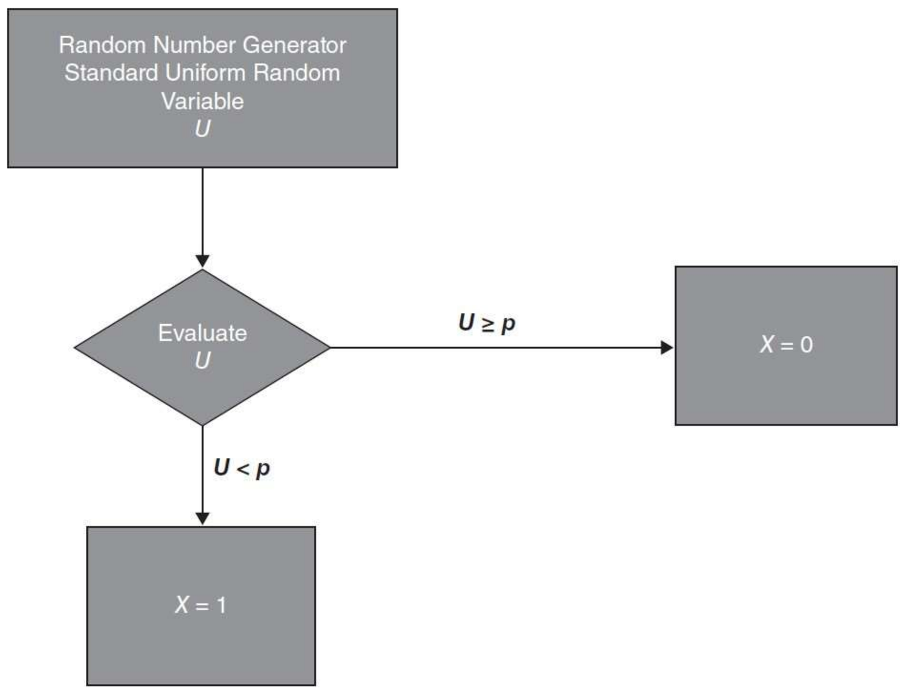
5. Evaluating integral, variance is

6. $\frac{(b2-b1)^2}{12}$

Do not need to memorize

Application of Uniform Random Variable: Monte Carlo simulation

- In a computer simulation, one way to model a Bernoulli variable is to start with a standard uniform variable, which determines when $b_1 = 0$ and $b_2 = 1$.



Normal Distribution

Week 3
Pre-class Prep 7/11

7. NORMAL

1. Normal distribution is also referred to as the Gaussian distribution.

2. For a random variable X, probability density function for the normal distribution is

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

$$X \sim N(\mu,\sigma^2),$$

which means "X is normally distributed with mean μ and variance σ^2."

Do not need to memorize

Properties of Normal Random Variable

1. Any linear combination of *independent* normal random variables is also normal.

$$Z = aX + bY \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2\right).$$

2. Example: If the log returns of individual stocks are independent and normally distributed, then the average return of those stocks will also be normally distributed.

3. The bell shape curve has 0 skewness and kurtosis of 3.

Recap: Standard Normal Random Variable

1. Standard normal distribution is $N(0, 1)$ with the probability density function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

2. Because a linear combination of independent normal distributions is also normal, standard normal distributions are the building blocks of many financial models.

3. To get a normal variable with a standard deviation of σ and a mean of μ, we simply multiply the standard normal variable ϕ by σ and add μ.

$$X = \mu + \sigma\varphi \sim N(\mu, \sigma^2).$$

Application 1: Standard Normal as "Building Blocks"

- To create two correlated normal variables, we can combine three independent standard normal variables, X_1, X_2, and X_3, as follows:

$$X_A = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_2$$

$$X_B = \sqrt{\rho}X_1 + \sqrt{1-\rho}X_3$$

- **Exercise**: Show that X_A and X_B are standard normal random variables.

X_A and X_B are standard normal r.v if they have mean 0 and variance 1
Can you show they have mean 0 and variance 1?

- **Exercise**: Show that X_A and X_B are indeed correlated with correlation coefficient ρ.

Exercise: Show that X_A and X_B are standard normal random variables

1. $E(X_A) = \sqrt{p}E(X_1) + \sqrt{(1-p)}E(X_2)$ = 0 (since both E(.) are 0)

2. $Var(X_A) = \sqrt{p}^2 Var(X_1) + \sqrt{1-p}^2 Var(X_2)$
$= p(1) + (1-p)(1)$
$= 1$

3. Therefore, X_A is a standard normal variable

4. Same reasoning applies to X_B by symmetry

Exercise: Show that X_A and X_B are correlated with correlation coefficient ρ

1. $\boldsymbol{Cov(X_A, XB) = E(X_A X_B) - E(X_A)E(X_B)}$

2. $= E\big((\sqrt{p}\mathrm{X}_1 + \sqrt{(1-p)}X_2) \times (\sqrt{p}X_1 + \sqrt{(1-p)}X_3\big) - E\left(\sqrt{p}X_1 + \sqrt{(1-p)}X_2\right) E\left(\sqrt{p}X_1 + \sqrt{(1-p)}X_3\right)$

3. = p

4. Because X_1, X_2 and X_3 are independent, many cross terms above, such as $E(X_1X_2)$ or $E(X_2X_3)$ evaluate to 0

5. Lastly, Correlation = Covariance / (std(X_A)std(X_B)). What are the respective standard deviations?

Log Return and Normal Distribution

1. Normal distributions are used throughout finance and risk management.

2. Normally distributed log returns are widely used in financial simulations as well, and form the basis of a number of financial models, including the Black-Scholes option pricing model.

3. One attribute that makes log returns particularly attractive is that they can be modeled using normal distributions.
    - A normal random variable can realize values ranging from $-\infty$ to ∞.
    - Simple return has a minimum, -100%.
    - But log return does not have a minimum as it can potentially be $-\infty$, and thus more amenable to modeling by a normal distribution..

Normal Distribution Confidence Intervals

	One-Tailed	Two-Tailed
1.00%	-2.33	-2.58
2.50%	-1.96	-2.24
5.00%	-1.64	-1.96
10.00%	-1.28	-1.64
90.00%	1.28	1.64
95.00%	1.64	1.96
97.50%	1.96	2.24
99.00%	2.33	2.58

- The normal distribution is symmetrical, 5% of the values are less than 1.64 standard deviations below the mean.
- The two-tailed value tells you that 95% of the mass is within ± 1.96 standard deviations of the mean. So, 95 % of the outcomes are less than -1.96 standard deviations from the mean.

Log Normal Distribution

Week 3
Pre-class Prep 8/11

8. LOGNORMAL [BACKGROUND MATERIAL]

1. Can we use another distribution and model the standard simple returns directly?

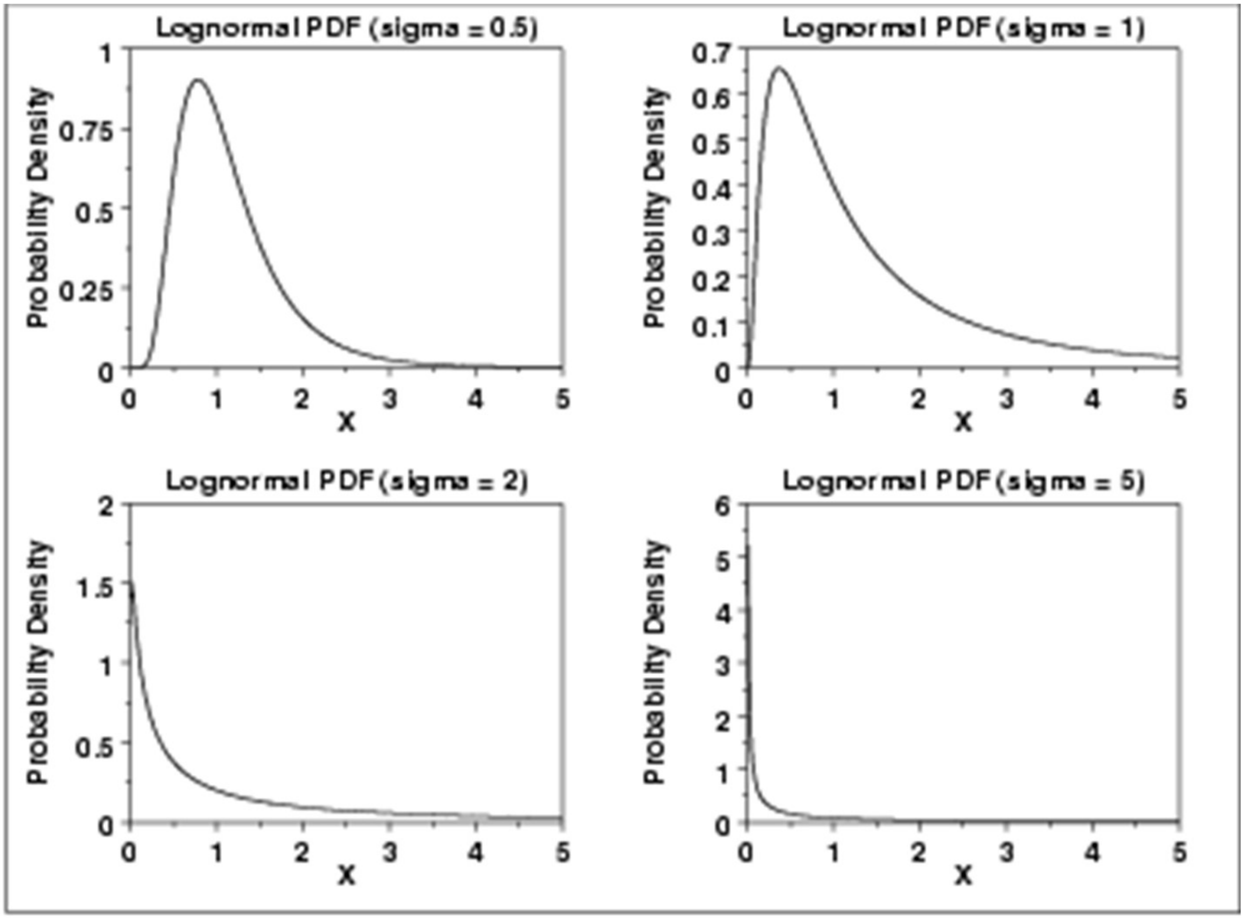
2. Yes, use the lognormal distribution instead. For $x > 0$,

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

3. If a variable has a lognormal distribution, then the log of that variable has a normal distribution. So, if log returns are assumed to be normally distributed, then one plus the simple return will be lognormally distributed

4. Using the lognormal distribution provides an easy way to ensure that returns less than -100% are avoided.

Not examinable



Not examinable

More on Lognormal Distribution

1. The lognormal pdf reproduced:

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

2. Given μ and σ , the mean and variance are given by

$$\mathbb{E}\left(X\right) = e^{\mu+\frac{1}{2}\sigma^2}$$
$$\mathbb{V}\left(X\right) = \left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}.$$

3. However it comes to modeling, it is often easier to work with log returns and normal distributions than with standard returns and lognormal distributions.

Chi-Squared Distribution

Week 3
Pre-class Prep 9/11

9. CHI-SQUARED

- If we have k independent standard normal variables, $Z_1, Z_2, \ldots, Z_k$, then the sum of their squares, S, has a chi-squared distribution.

$$S := \sum_{i=1}^{k} Z_i^2; \qquad S \sim \chi_k^2$$

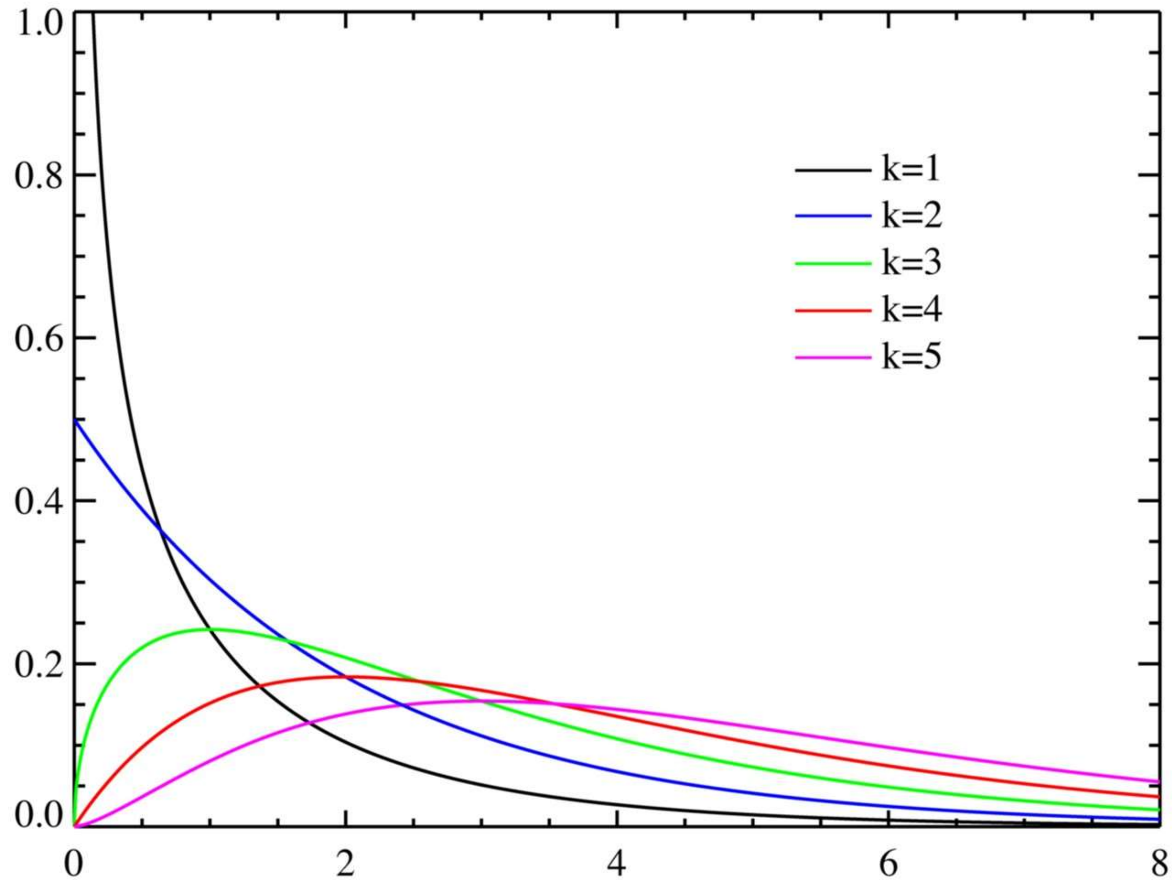
- The variable k is commonly referred to as the degrees of freedom.
- For positive values of x, the probability density function for the chi-squared distribution is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}},$$

Do not need to memorize formula

where Γ is the gamma function: $\Gamma(n) := \int_0^\infty x^{n-1} e^{-x}\, dx.$

Chi-Square Probability Density Function



Application: Hypothesis testing for variance estimate

- Recap: If we have k independent standard normal variables, Z_1, $Z_2, \ldots, Z_k$, then the **<u>sum of their squares</u>**, S, has a chi-squared distribution.
- The variance estimate ("sampling distribution of the sample variance" of a normal variable therefore will have a chi-squared distribution

Application: Sample Variance and Chi-Square Statistic

- Let $Z_1, Z_2, \ldots, Z_k$ be k independent random variables, and each Z is a standard normal random variable. The sum of the squared Zs also follows a chi-squared distribution with k degrees of freedom.

$$\sum_{i=1}^{k} Z_i^2 \overset{d}{\sim} \chi_k^2.$$

- The unbiased sample variance s^2 is a random variable owing to random sampling.

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{\sigma^2} = (n-1) \frac{s^2}{\sigma^2} \overset{d}{\sim} \chi_{n-1}^2.$$

Do not need to memorize formula, but you must <u>remember that variance is chi-sq distributed!</u>

More on Chi-Squared Distribution

- The chi-squared distribution is widely used in risk management, and in statistics in general, for hypothesis testing.

- The mean of the distribution is k, and the variance is $2k$.

- As k increases, the chi-squared distribution becomes increasingly symmetrical. As k approaches infinity, the chi-squared distribution converges to the normal distribution.

☆ t -Distribution

Week 3
Pre-class Prep 10/11

10. STUDENT'S T

- **William Sealy Gosset**, an employee at the Guinness brewery in Dublin, used the pseudonym Student for journal publications.
- If *Z* is a standard normal variable and *U* is a chi-square variable with *k* degrees of freedom, which is independent of *Z*, then the random variable *X*, follows a t distribution with k degrees of freedom.

$$X = \frac{Z}{\sqrt{U/k}},$$

No need to memorize formula, but please <u>remember that "normal/chi-sq is t-distributed"</u>

- The probability density function of Student's t random variable is

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

where k is the degrees of freedom and $\Gamma(x)$ is the gamma function.

Application: T-distribution

- We note that the t-distributed random variable will be the "test statistic of sample mean estimates"

Recall also we take this ratio in order to derive the "standardized" variable. i.e. standardized with respect to variance

$$X = \frac{\boxed{Z}}{\sqrt{\boxed{U}/k}},$$

Distribution of sample mean estimate is normal (linear combination of normal variables)

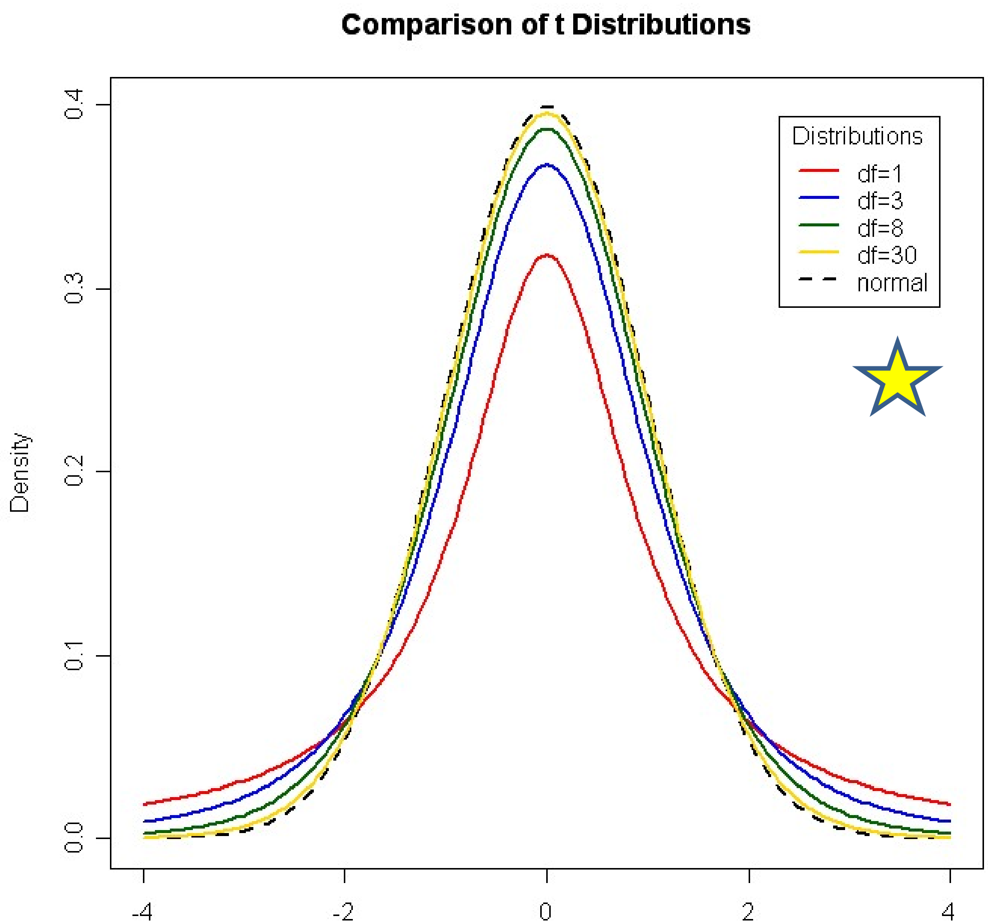
Variance of sample mean estimate is chi-sq (linear combination of square of normal variables)

This is why most research papers look at statistical significance of the "t-statistic"

Characteristics of Student's t Distribution

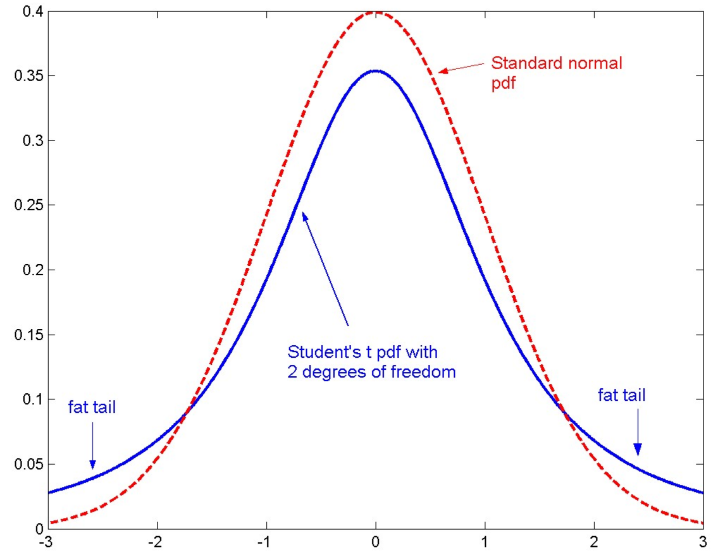
- The t distribution is symmetrical around its mean, which is equal to zero.
- For low values of k, the t distribution looks very similar to a standard normal distribution, except that it displays excess kurtosis.
- As k increases, this excess kurtosis decreases. In fact, as k approaches infinity, the t distribution converges to a standard normal distribution.
- The variance of the t distribution for $k > 2$ is $\frac{k}{k-2}$. As k increases, the variance of the t distribution converges to one, the variance of the standard normal distribution.

Comparison of t Distributions



Probability Density Function of *t* Statistics

- The pdf of *t* statistics have “fatter” tails compared to standard normal.
- When the number of degrees of freedom is ∞, Student’s *t* pdf becomes standard normal pdf.



Chi-Square and Student *t* Distributions

- If random variable $Z \overset{d}{\sim} N(0,1)$, then

$$V := Z^2 \overset{d}{\sim} \chi_1^2.$$

- More generally,

$$\sum_{i=1}^{n} Z_i^2 \overset{d}{\sim} \chi_n^2.$$

No need to memorize formula, but reiterating conceptually <u>that "normal/chi-sq is t-distributed"</u>

- If $X \overset{d}{\sim} N(0,1)$ and both *X,V* $V \overset{d}{\sim} \chi_n^2$ are independent, then the following statistic follows Student's t distribution with n degrees of freedom, i.e.,

$$\frac{X}{\sqrt{\frac{V}{n}}} \overset{d}{\sim} t_n.$$

F -Distribution

Week 3
Pre-class Prep 11/11

11. F-DISTRIBUTION

- If U_1 and U_2 are two independent chi-squared distributions with k_1 and k_2 degrees of freedom, respectively, then X,

$$X = \frac{U_1/k_1}{U_2/k_2} \sim F(k_1, k_2)$$

  follows an F-distribution with parameters k_1 and k_2.

- The probability density function of the F-distribution is

$$f(x) = \frac{\sqrt{\frac{(k_1 x)^{k_1} k_2^{k_2}}{(k_1 x + k_2)^{k_1+k_2}}}}{x B(k_1/2, k_2/2)},$$

  where $B(x,y)$ is the beta function:

$$B(x,y) = \int_0^1 z^{x-1}(1-z)^{y-1}\,dz.$$

No need to memorize any formulas here, but remember <u>that "chi-sq/chi-sq is F-distributed"</u>

A Bit More on F-Distribution

- The mean and variance of the F-distribution are as follows

$$\mu = \frac{k_2}{k_2 - 2}, \qquad \text{for } k_2 > 2,$$
$$\sigma^2 = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}, \qquad \text{for } k_2 > 4.$$

Not examinable; most of our use for F-distributions will be in code

- As k_1 and k_2 increase to infinity, the mean and mode converge to one, and the F-distribution converges to a normal distribution.

- The square of a variable with a t distribution has an F-distribution. More specifically, if X is a random variable with a t distribution with k degrees of freedom, then X^2 has an F-distribution with 1 and k degrees of freedom:

$$X^2 \sim F(1, k).$$

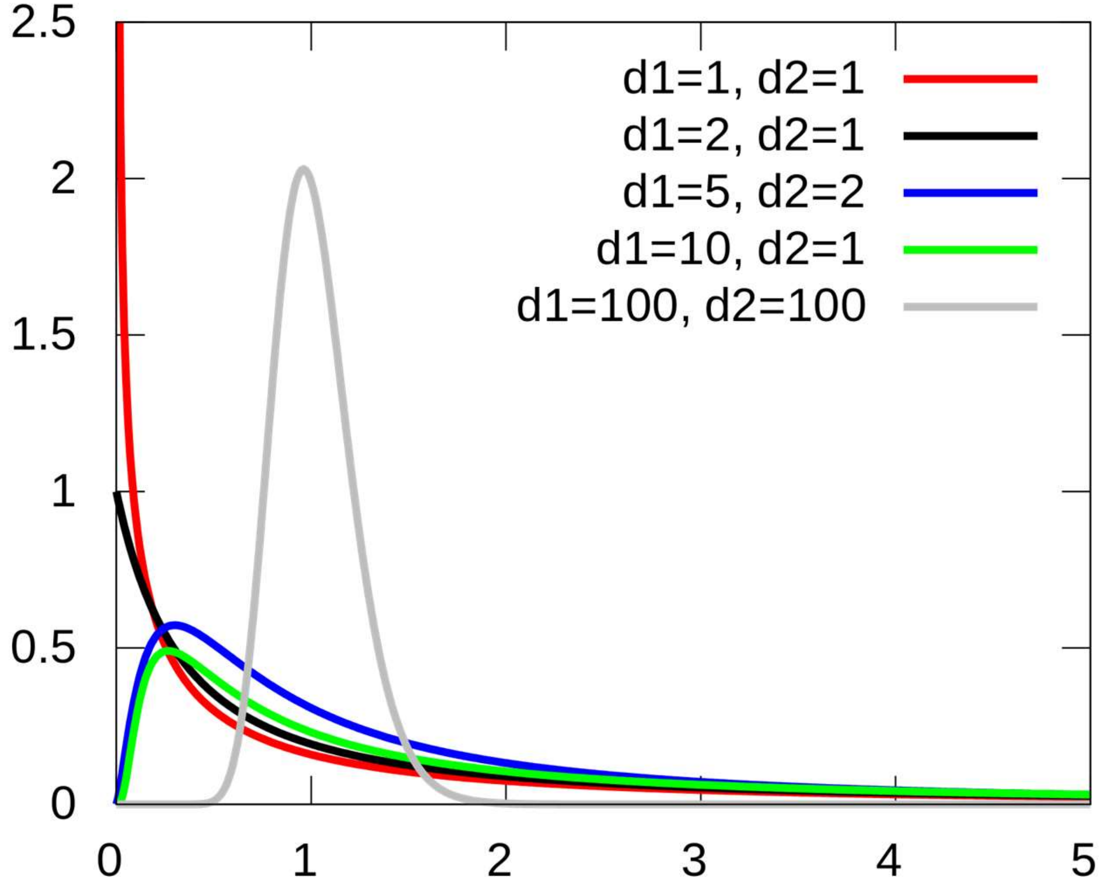
Application 1: F test (overview only)

- Consider 2 models, where model 1 is ‘nested’ within model 2.
- This usually means that model 2 has more parameters than model 1. Also, model 2 has all the parameters that model 1 has, but not vice versa
- We want to test if model 2 with more parameters gives a significantly better fit to the data than model 1. One approach to this problem is to use an F-test
- If there are n data points, then we can calculate the F statistic in the following way. The statistic will follow an F distribution

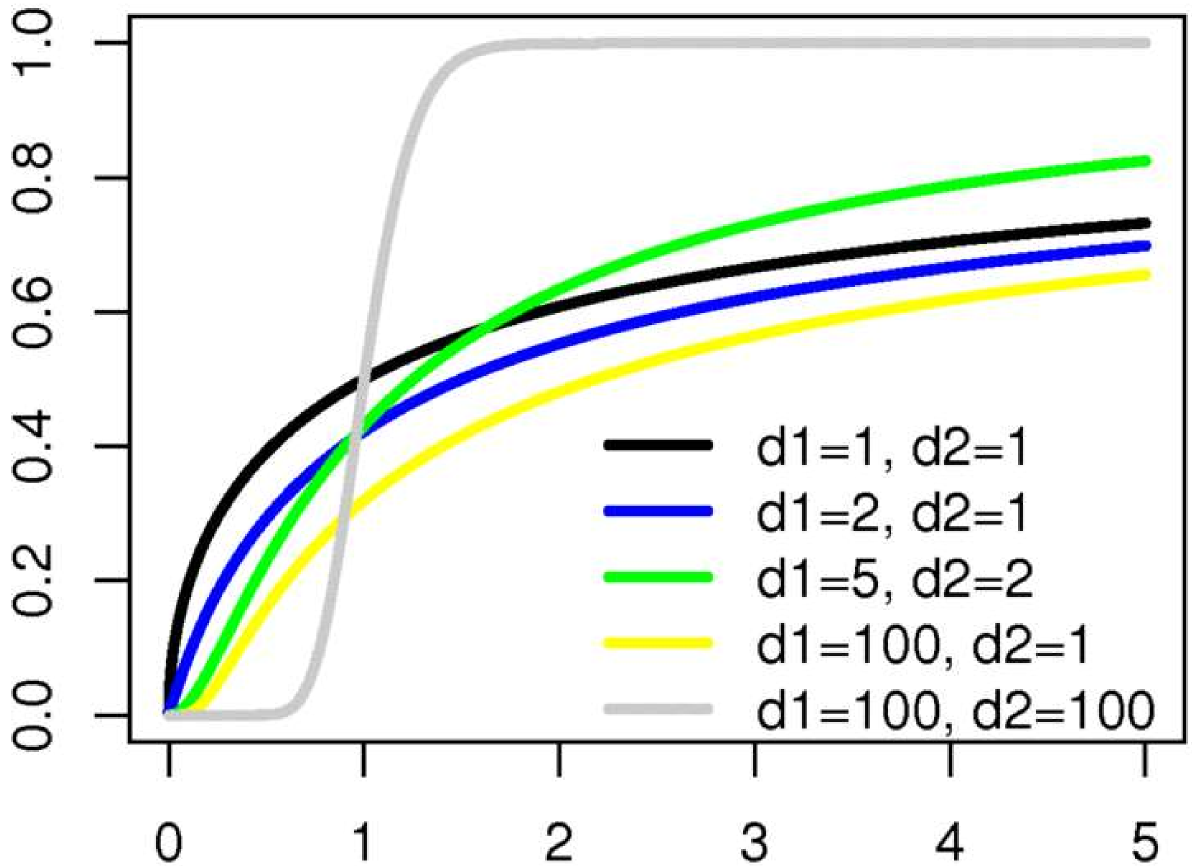
$$F = \frac{\left(\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{p_2 - p_1}\right)}{\left(\frac{\mathrm{RSS}_2}{n - p_2}\right)},$$

We will revisit this test in future classes. This is just a rough overview

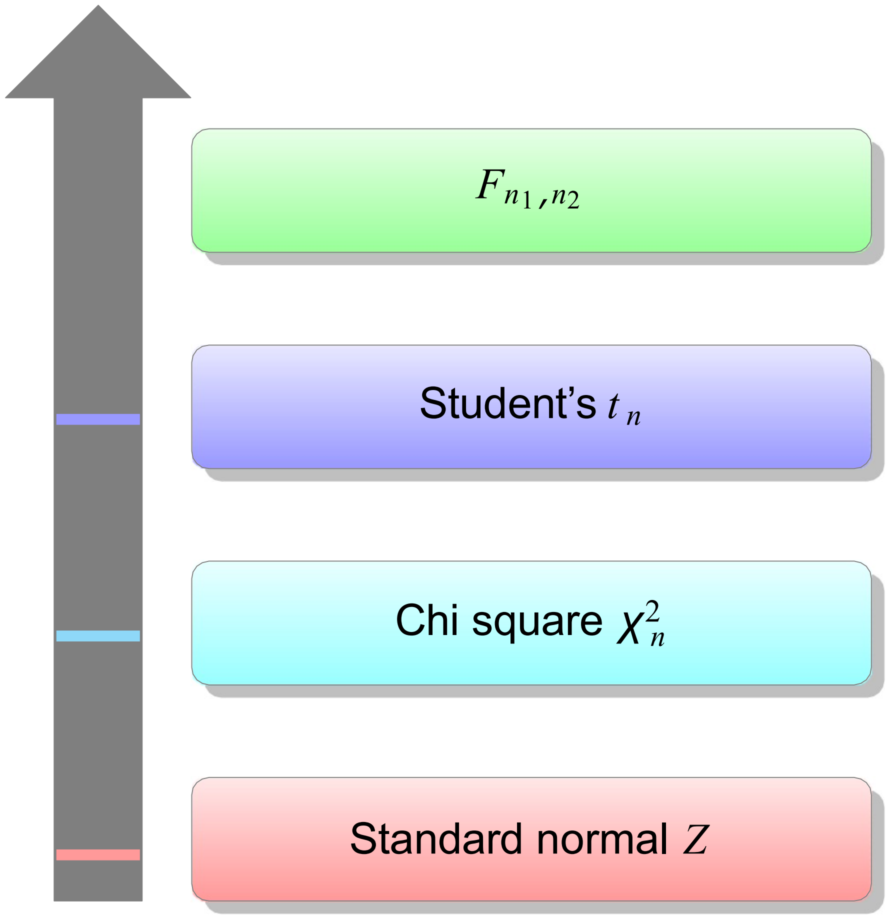
F Probability Density Function



CDF of F Distributions



Summary



$$F_{n_1,n_2} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$$

$$\lim_{n\to\infty} t_n \longrightarrow Z$$

$$V := \sum_{i=1}^{n} Z_i^2 \overset{d}{\sim} \chi^2_n$$

$$Z = \frac{r-\mu}{\sigma}$$

Takeaways

- All the 10 distributions are parametric, being dependent on parameters that can be interpreted intuitively
- χ^2 is the sum of k squared of independent standard normal variables
- t variable is made of a standard normal random variable and a χ^2 random variable, which are independent.
- The F random variable is the ratio of two independent χ^2 random variables.