37262 Mathematical Statistics

Lecture 4

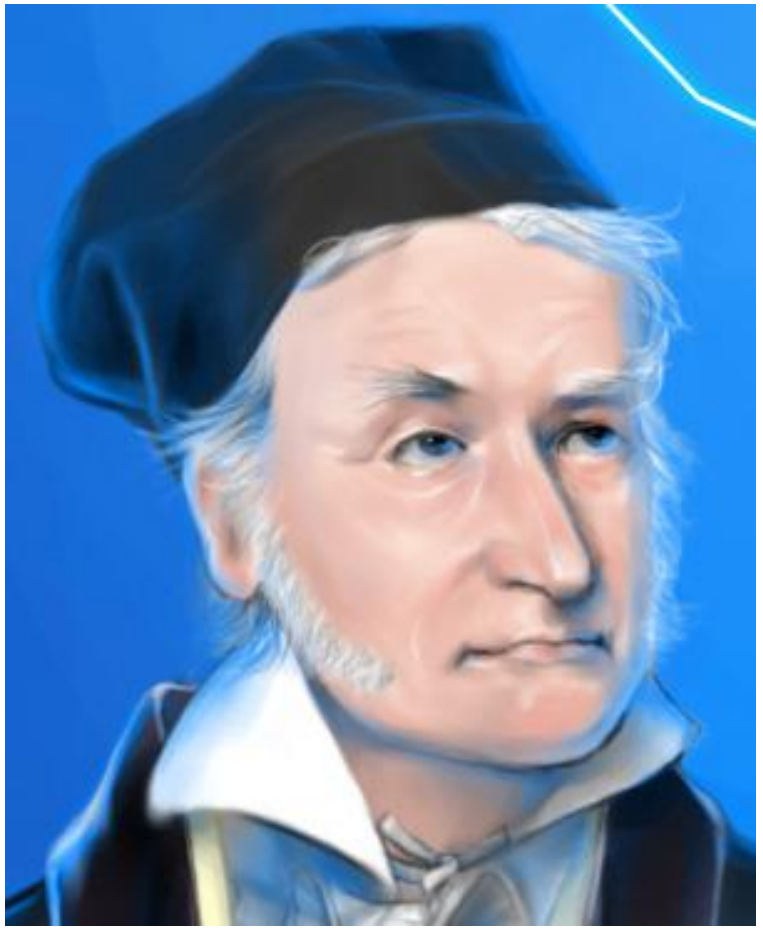
Normal Distribution

- The most commonly seen variables used in statistics are **Normal** or **Gaussian** variables.

- This is the classic “bell curve” shape.

- If the variable $Z \sim N(\mu,\sigma^2)$ then, for $\sigma^2 > 0$, then the density function of Z is

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(z-\mu)^2}{2\sigma^2}\right)} \quad \text{for } -\infty < z < \infty.$$



Karl Friedrich Gauss

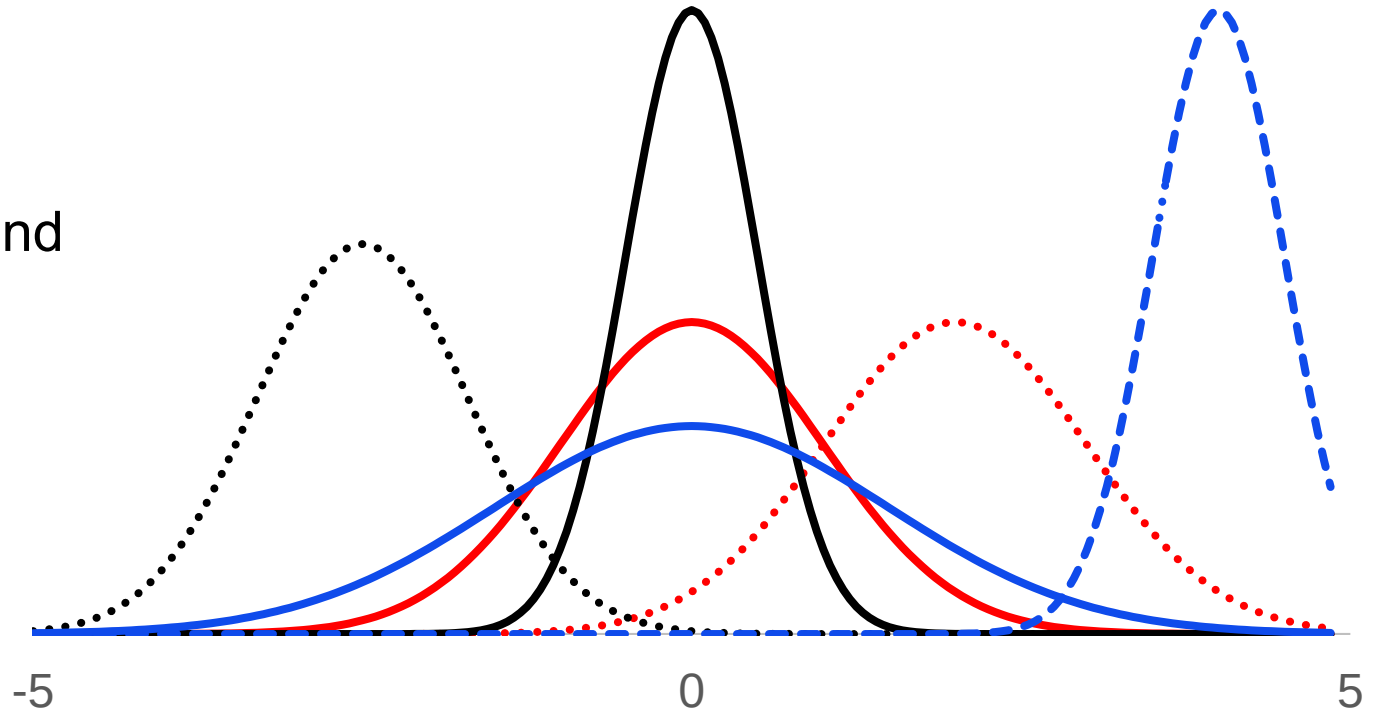
(*1777 – 1855*)

Normal Distribution

- Although always “bell curve” shaped and symmetric about its mean value, the exact shape and position of the curve depends on two parameters.

- A larger μ shifts the centre of the curve upwards. A larger σ^2 increases the variance or the spread of the observations, so widens the curve.

- $Z \sim N(0,1)$ is often called a **standard normal** variable.



Central Limit Theorem

- The main reason that the normal distribution is so important to many statistical problems is the **central limit theorem**.

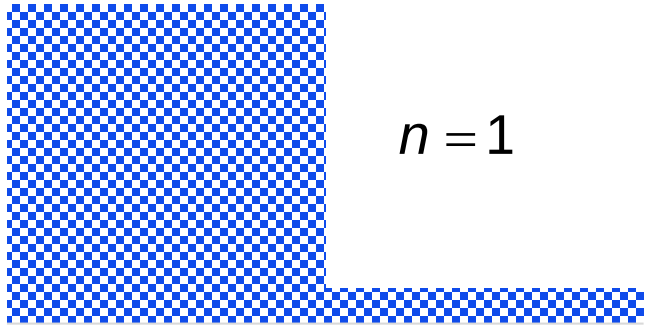
- The central limit theorem states that, if $X_1, X_2, ..., X_n$ are independent variables, each drawn from the same distribution with mean $\mu < \infty$ and variance $\sigma^2 < \infty$, then the variable

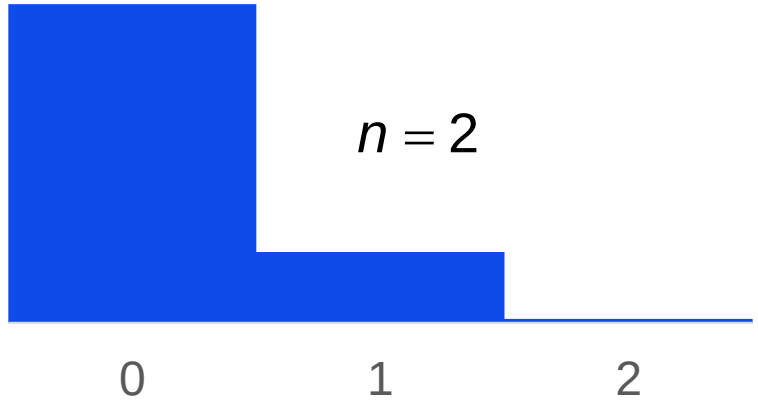
$$Z = \lim_{n\to\infty} \left(\frac{\bar{X}_n - \mu}{\left(\frac{\sigma}{\sqrt{n}} \right)} \right) \sim N(0,1) \qquad \text{where } \bar{X}_n \text{ is the sample mean } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- In other words, whatever distribution (with finite mean and variance) samples are drawn from, eventually the quantity $\sqrt{n}(\bar{X}_n - \mu)$ is asymptotically normally distributed for large *n*.

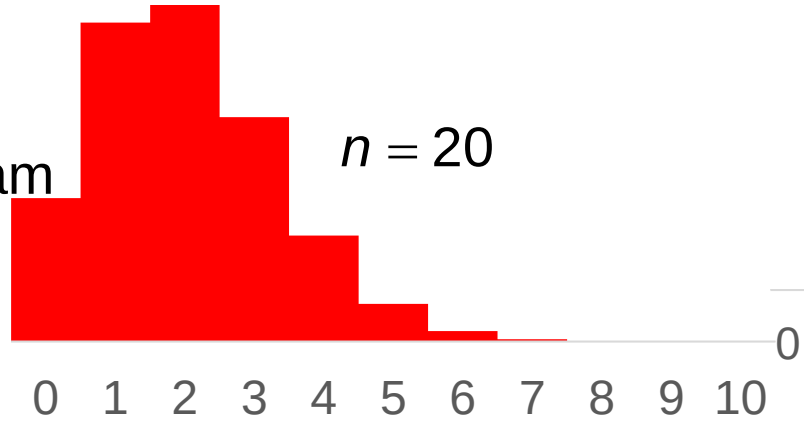
Central Limit Theorem

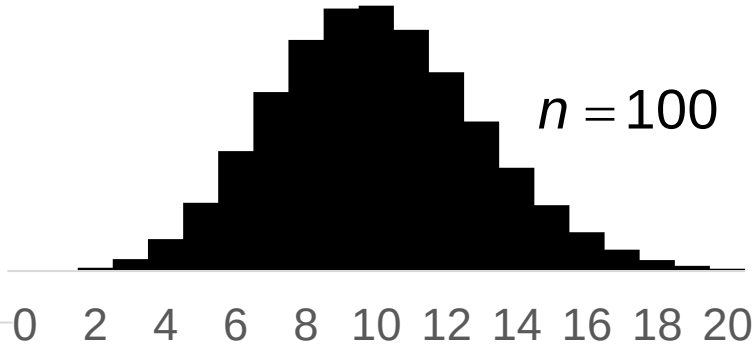
- A formal proof of the central limit theorem is beyond the scope of this subjecy, but we can see a few examples of why it works for some distributions we have already seen.





- Consider n independent variables $T_i \sim Bern(0.1)$. We already know that $\sum_{i=1}^{n} T_i \sim Bin(n, 0.1)$.

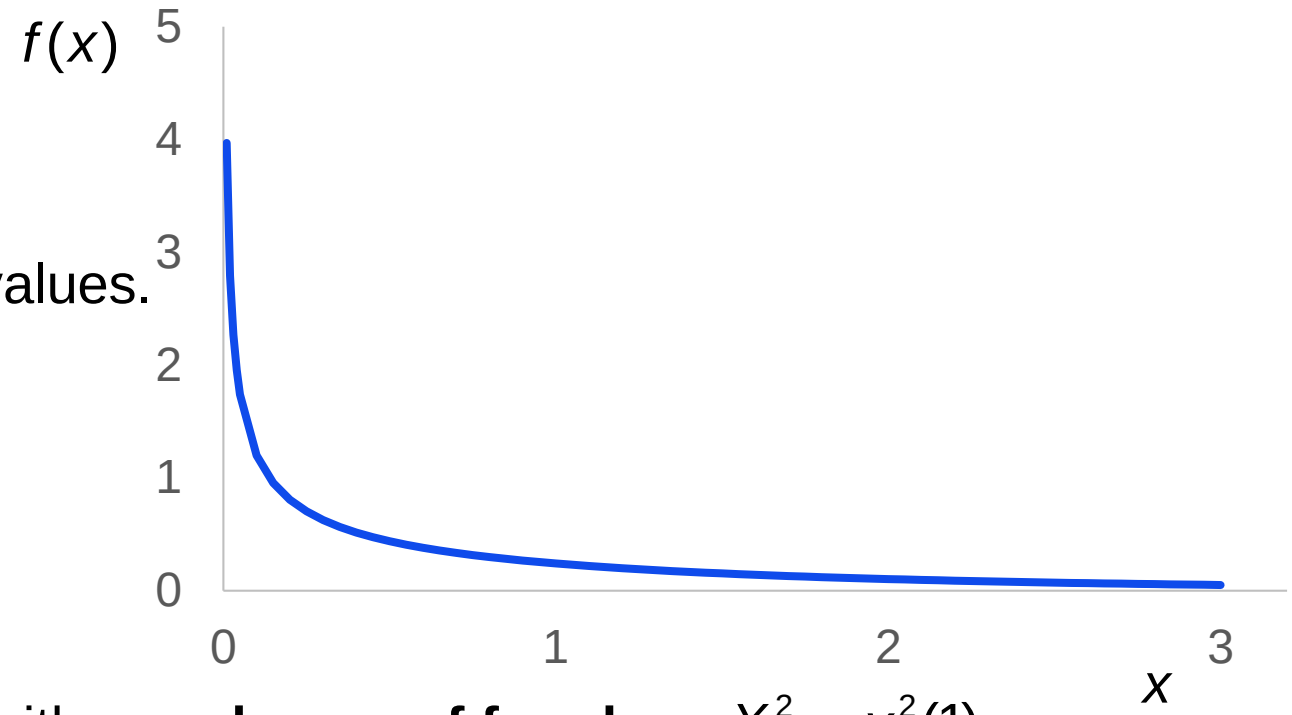




- For large n, the binomial histogram closely resembles a normal curve, centred around its expected value of $0.1n$.

Functions of Normal Variables

- Consider the case where we have a standard normal variable $X \sim N(0,1)$.
- We consider the distribution of X^2.
- Clearly X^2 can only take non-negative values.
- Because the mode of X is at zero, the mode of X^2 is also at zero.



- The distribution of X^2 is **chi-squared** with **one degree of freedom** $X^2 \sim \chi^2(1)$

Chi-Squared Distribution

- We can now calculate the probability density function of $X^2 = Z \sim \chi^2(1)$.

- The density function of X is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ for $-\infty < x < \infty$.

- Our transformation $Z = X^2$ is not one-to-one, since, for example

  $X = -1$ and $X = 1$ both give $Z = 1$.

- As the distribution is symmetric about 0, we can instead only consider positive values of X and then double this up for the negative half of the distribution.

Chi-Squared Distribution

- Our change of variable formula for one variable is $f_Z(z) = f_X(x(z))\dfrac{dx}{dz}$
- $z = x^2$ hence $x = \sqrt{z}$ giving $\dfrac{dx}{dz} = \dfrac{1}{2\sqrt{z}}$.
- This gives $f_Z(z) = 2\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z}{2}}\left(\dfrac{1}{2\sqrt{z}}\right)$ for $0 \le z < \infty$.
- The density function of $Z \sim \chi^2(1)$ is therefore $f_Z(z) = \begin{cases} \dfrac{1}{\sqrt{2\pi z}}e^{-\frac{z}{2}} & 0 \le z < \infty \\ 0 & \text{otherwise} \end{cases}$

Chi-Squared Distribution

- We can now show that the sum of the squares of *n* independent standard normal variables has a **chi-squared distribution** with *n* degrees of freedom.

- That is, if $X_1 \sim N(0,1), X_2 \sim N(0,1), \ldots, X_n \sim N(0,1)$ are independent, then $S = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$.

- We will justify that the density function of S is $f_S(s) = \begin{cases} \dfrac{1}{\sqrt{2^n}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} & 0 \le s < \infty \\ 0 & \text{otherwise} \end{cases}$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy = (\alpha - 1)!$ when α is a positive integer.

Chi-Squared Distribution

- We can already justify that $f_S(s) = \begin{cases} \dfrac{1}{\sqrt{2^n}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} & 0 \le s < \infty \\ 0 & \text{otherwise} \end{cases}$ for $n = 1$ since

  setting $n = 1$ simplifies this to the $\chi^2(1)$ density we have already seen (noting that $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$.)

- We now need to justify that the sum of $S = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$ and $T = \sum_{i=n-1}^{n+m} X_i^2 \sim \chi^2(m)$ is itself chi-squared distributed with $m + n$ degrees of freedom. $(S + T) \sim \chi^2(m + n)$.

- By induction, this will justify that this holds for all positive integers *m* and *n*.

Chi-Squared Distribution

- Let $S \sim \chi^2(n)$ and $T \sim \chi^2(m)$ be independent random variables.
- Their joint density function is then $f_{S,T}(s,t) = \begin{cases} \dfrac{1}{\sqrt{2^n}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} \dfrac{1}{\sqrt{2^m}\Gamma\left(\dfrac{m}{2}\right)} e^{-\frac{t}{2}} t^{\frac{m}{2}-1} & (s,t) \in [0,\infty)^2 \\ 0 & \text{otherwise} \end{cases}$
- Setting a change of variables $U = S + T$ and $V = \dfrac{S}{S+T}$ we obtain $S = UV$ and $T = U(1-V)$.
- The Jacobian associated with this change of variables is therefore

$$\boldsymbol{J} = \begin{pmatrix} \dfrac{\partial S}{\partial U} & \dfrac{\partial S}{\partial V} \\ \dfrac{\partial T}{\partial U} & \dfrac{\partial T}{\partial V} \end{pmatrix} = \begin{pmatrix} V & U \\ (1-V) & -U \end{pmatrix} \text{ hence } \left|\det(\boldsymbol{J})\right| = U.$$

Chi-Squared Distribution

- Setting $S = UV$ and $T = U(1-V)$ and $\left|\det(\boldsymbol{J})\right| = U$ into the joint density of S and T,

$$f_{S,T}(s,t) = \begin{cases} \dfrac{1}{\sqrt{2^n}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} \dfrac{1}{\sqrt{2^m}\Gamma\left(\dfrac{m}{2}\right)} e^{-\frac{t}{2}} t^{\frac{m}{2}-1} & (s,t) \in [0,\infty)^2 \\ 0 & \text{otherwise} \end{cases}$$

we obtain $$f_{U,V}(u,v) = \begin{cases} \dfrac{1}{\sqrt{2^{m+n}}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{u}{2}} (uv)^{\frac{n}{2}-1} \dfrac{1}{\Gamma\left(\dfrac{m}{2}\right)} (u(1-v))^{\frac{m}{2}-1} u & (u,v) \in [0,\infty) \times [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Chi-Squared Distribution

- This then simplifies to

$$f_{U,V}(u,v)=\begin{cases}\dfrac{1}{\sqrt{2^{m+n}}\Gamma\left(\dfrac{m+n}{2}\right)}e^{-\frac{u}{2}}u^{\frac{m+n}{2}-1}\dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}v^{\frac{n}{2}-1}(1-v)^{\frac{m}{2}-1} & (u,v)\in[0,\infty)\times[0,1] \\ 0 & \text{otherwise}\end{cases}$$

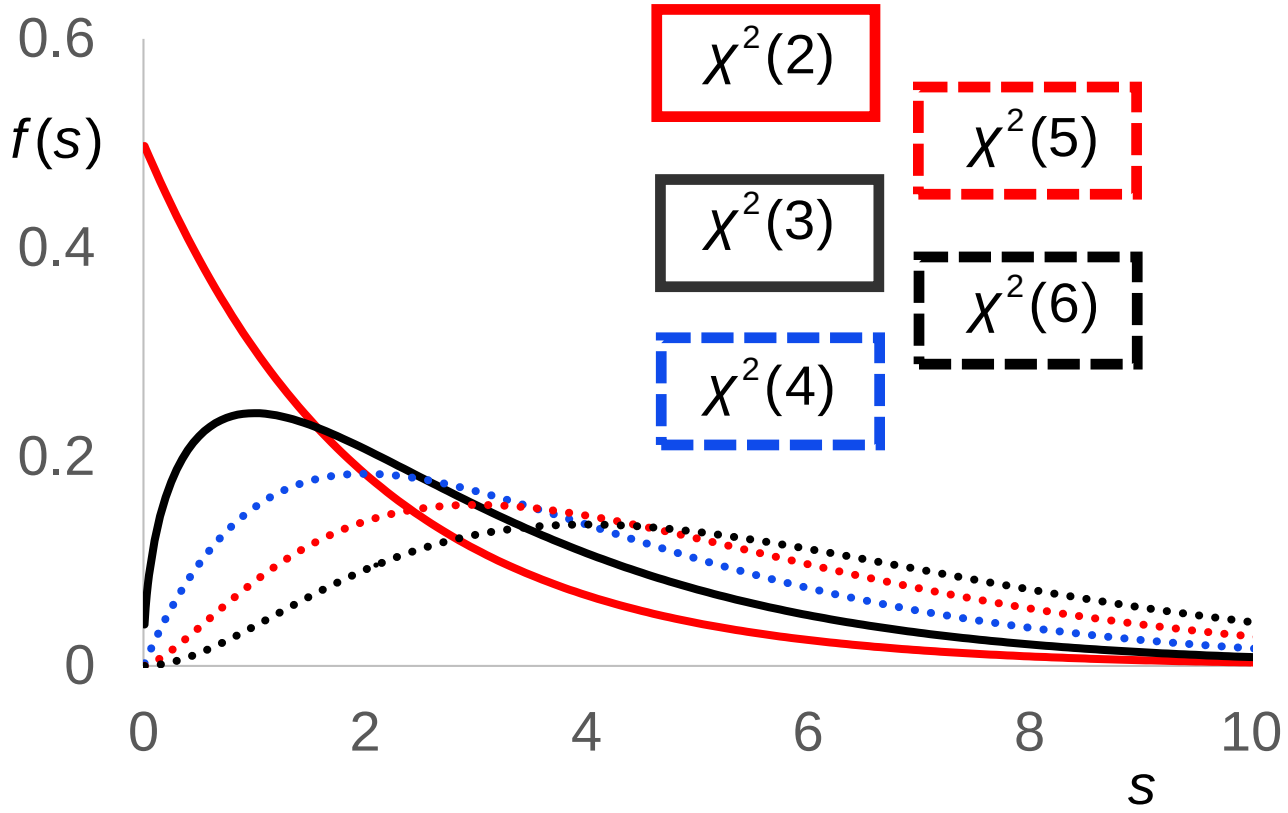
- This separates to give two independent variables $U \sim \chi^2(m+n)$ and $V \sim beta\left(\dfrac{n}{2},\dfrac{m}{2}\right)$.

Chi-Squared Distribution

- The density function of $S \sim \chi^2(n)$

$$f_S(s) = \begin{cases} \dfrac{1}{\sqrt{2^n}\,\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} & 0 \le s < \infty \\ 0 & \text{otherwise} \end{cases}$$

has a very different shape for different degrees of freedom.



- For $n = 1$ or 2, the mode is at 0. For more degrees of freedom, the mode is at $n - 2$.
- Since $S \sim \chi^2(n)$ can be considered as the sum of n independent $N(0,1)$ variables, $X_i \sim N(0,1)$

$$E(S) = E\left(\sum_{i=1}^{n} X_i^{\,2}\right) = nE(X_i^{\,2}) = n.$$

Chi-Squared Goodness of Fit Tests

- You may have seen a chi-squared variable before in the context of evaluating how well a proposed model fits a given dataset.

- This relies on both the central limit theorem and the fact that the square of n independent standard normal variables is a $\chi^2(n)$ variable.

- Consider the simple example of flipping a coin which we believe to be biased towards Heads. We have been told that it will land Heads each flip with probability 0.7.

- We flip it 40 times and observe 20 Heads and 20 Tails. Should be reject the stated hypothesis that it lands Heads with probability 0.7?

Chi-Squared Goodness of Fit Tests

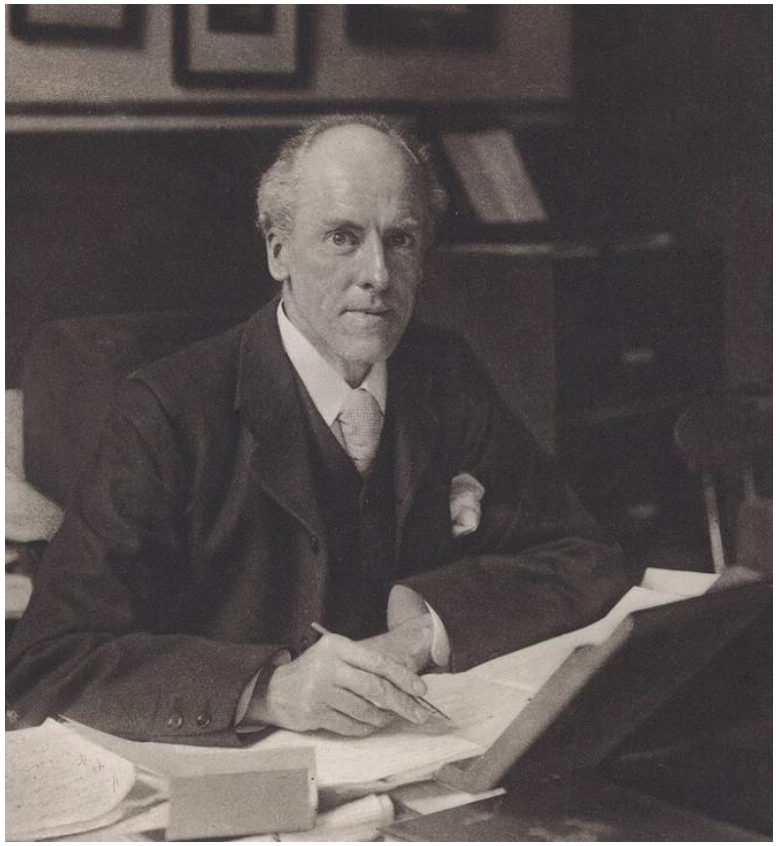
- In the most general case, if we hypothesise that a proportion p out of N flips will land Heads and we observe m Heads from the experiment, the central limit theorem tell us that the test statistic $\frac{m-Np}{\sqrt{Np(1-p)}}$ is approximately a standard normal.
- Note, for a binomial variable $X \sim Bin(N,p)$, $E(X)=np$ and $\text{Var}(X)=np(1-p)$.
- For $\frac{m-Np}{\sqrt{Np(1-p)}} \sim N(0,1)$ we have that $\left[\frac{m-Np}{\sqrt{Np(1-p)}}\right]^2 \sim \chi^2(1)$
- Expanding this gives $\left[\frac{m-Np}{\sqrt{Np(1-p)}}\right]^2 = \frac{(m-Np)^2}{Np(1-p)} = \frac{(m-Np)^2}{Np} + \frac{(N-m-N(1-p))^2}{N(1-p)}$

Chi-Squared Goodness of Fit Tests

- $\frac{(m-Np)^2}{Np} + \frac{(N-m-N(1-p))^2}{N(1-p)}$ is of the form

  $\sum \frac{(O_i - E_i)^2}{E_i}$ where O_i is the observed count in outcome i and E_i is its expected count under the null model.

- This is sometimes called the **Pearson statistic**, after Karl Pearson.



Karl Pearson
(*1857 – 1936*)

- Provided the central limit approximation holds (usually assuming all expected counts are at least 5) the Pearson statistic is approximately chi-squared distributed.

Chi-Squared Goodness of Fit Tests

- For our example of 40 coin flips, under the null model of Heads with probability 0.7 $E_{Heads} = 28$ and $E_{Tails} = 12$.
- The observed counts were $O_{Heads} = 20$ and $O_{Tails} = 20$.
- Our Pearson statistic is therefore $\frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} \approx 7.62$.
- Comparing this with a chi-square distribution with one degree of freedom, we see that if $Y \sim \chi^2(1)$, then $P(Y < 7.62) \approx 0.994$.
- If testing with significance level 0.05 (or even 0.01) we would reject the null hypothesis.

Chi-Squared Goodness of Fit Tests

- Although a little harder to prove, the Pearson statistic extends to multi-category goodness of fit, again summing $\frac{(O_i - E_i)^2}{E_i}$ over all possible cells with expectations calculated under the null hypothesis.
- In general, if the null hypothesis proposes n proportions which the dataset should fit, we have $n - 1$ degrees of freedom (since the total number of observations is constrained and hence all outcomes not included in the first $n - 1$ categories are surely in the final category.

Chi-Squared Goodness of Fit Tests: Contingency Tables

- The other commonly seen null hypothesis for such datasets is that the row proportions and column proportions are independent. In this case, a table with r rows and c columns gives rise to a chi-squared variable with $(r-1)(c-1)$ degrees of freedom.

- Consider an insurance company which looks at claims, sorted by colour of car.

	Red	**Black**	**Blue**
No Claim	320	500	380
Claim	80	100	120

Observed

Chi-Squared Goodness of Fit Tests: Contingency Tables

- The proportions of each colour are $p_{\text{Red}} = \frac{400}{1500}$, $p_{\text{Black}} = \frac{600}{1500}$ and $p_{\text{Blue}} = \frac{500}{1500}$
- The proportion of claim categories are $p_{\text{No}} = \frac{1200}{1500}$ and $p_{\text{Claim}} = \frac{300}{1500}$.
- Under the null model, for example, $E_{\text{Red, No}} = 1500 p_{\text{Red}} p_{\text{No}} = 1500\left(\frac{400}{1500}\right)\left(\frac{1200}{1500}\right) = 320.$

	Red	**Black**	**Blue**
No Claim	320	500	380
Claim	80	100	120

Observed

Red	**Black**	**Blue**
320	480	400
80	120	100

Expected

Chi-Squared Goodness of Fit Tests: Contingency Tables

- Our Pearson statistic is then

$$\frac{(320-320)^2}{320}+\frac{(80-80)^2}{80}+\frac{(500-480)^2}{480}+\frac{(100-120)^2}{120}+\frac{(380-400)^2}{400}+\frac{(120-100)^2}{100}\approx 9.167$$

- Comparing this with $Y \sim \chi^2(2)$, then $P(Y < 9.167) \approx 0.990$.

- Again, with significance level 0.05 (or even 0.01) we would reject the null hypothesis.

	Red	**Black**	**Blue**
No Claim	320	500	380
Claim	80	100	120

Observed

Red	**Black**	**Blue**
320	480	400
80	120	100

Expected

Functions of Normal Variables

- Let us now consider two independent variables $X \sim N(0,1)$ and $Z \sim \chi^2(n)$.

- The joint density of these is therefore

$$f_{X,Z}(x,z) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{z}{2}} z^{\frac{n}{2}-1} & (x,z) \in (-\infty,\infty) \times [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

- We now consider the change of variables $Y = Z$ and $T = \frac{X}{\sqrt{\frac{Z}{n}}}$.

Functions of Normal Variables

- $Y = Z$ and $T = \dfrac{X}{\sqrt{\dfrac{Z}{n}}}$ gives $X = T\sqrt{\dfrac{Y}{n}}$ and $Z = Y$.
- The Jacobian is therefore $\boldsymbol{J} = \begin{pmatrix} \dfrac{\partial X}{\partial T} & \dfrac{\partial X}{\partial Y} \\ \dfrac{\partial Z}{\partial T} & \dfrac{\partial Z}{\partial Y} \end{pmatrix} = \begin{pmatrix} \sqrt{\dfrac{Y}{n}} & \dfrac{T}{2}\sqrt{\dfrac{1}{nY}} \\ 0 & 1 \end{pmatrix}$ hence $\left|\det(\boldsymbol{J})\right| = \sqrt{\dfrac{Y}{n}}$.

$$f_{T,Y}(t,y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{t^2 y}{2n}} \dfrac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} \sqrt{\dfrac{y}{n}} & (t,y) \in (-\infty,\infty) \times [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Functions of Normal Variables

- We can simply this to

$$f_{T,Y}(t,y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2 y}{2n}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} \sqrt{\frac{y}{n}} & (t,y) \in (-\infty,\infty)\times[0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

or

$$f_{T,Y}(t,y) = \begin{cases} \frac{1}{\sqrt{2\pi n}} e^{-\frac{y}{2}\left(1+\frac{t^2}{n}\right)} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n+1}{2}-1} & (t,y) \in (-\infty,\infty)\times[0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Functions of Normal Variables

- The marginal density $f_T(t) = \int_0^\infty f_{T,Y}(t,y)dy$
- $f_T(t) = \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} \int_0^\infty e^{-\frac{y}{2}\left(1+\frac{t^2}{n}\right)} y^{\frac{n+1}{2}-1} dy$
- With change of variables $v = \frac{y}{2}\left(1+\frac{t^2}{n}\right)$ hence $\frac{dv}{dy} = \frac{1}{2}\left(1+\frac{t^2}{n}\right)$

$$f_T(t) = \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} 2^{\frac{n+1}{2}} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \int_0^\infty e^{-v} v^{\frac{n+1}{2}-1} dv$$

Functions of Normal Variables

- We know that the gamma function is defined such that $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ hence

$$f_T(t) = \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} 2^{\frac{n+1}{2}} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \int_0^\infty e^{-v} v^{\frac{n+1}{2}-1} dv$$

$$= \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} 2^{\frac{n+1}{2}} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \Gamma\left(\frac{n+1}{2}\right)$$

- Tidying this up, we obtain $f_T(t) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$ for $-\infty < t < \infty$

t-Distribution

- A variable $T \sim t_n$ has a **t-distribution** or **Student's t-distribution** with ***n* degrees of freedom** and probability density function

$$f_T(t) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad \text{for } -\infty < t < \infty$$

- This was developed by William Sealy Gosset, who was studying quality control issues at the Guinness brewery in Ireland. He published his work under the name "Student."

- A realisation of a t-distribution can be simulated by taking the ratio of a standard normal variable to the square root of a $\chi^2(n)$ variable.

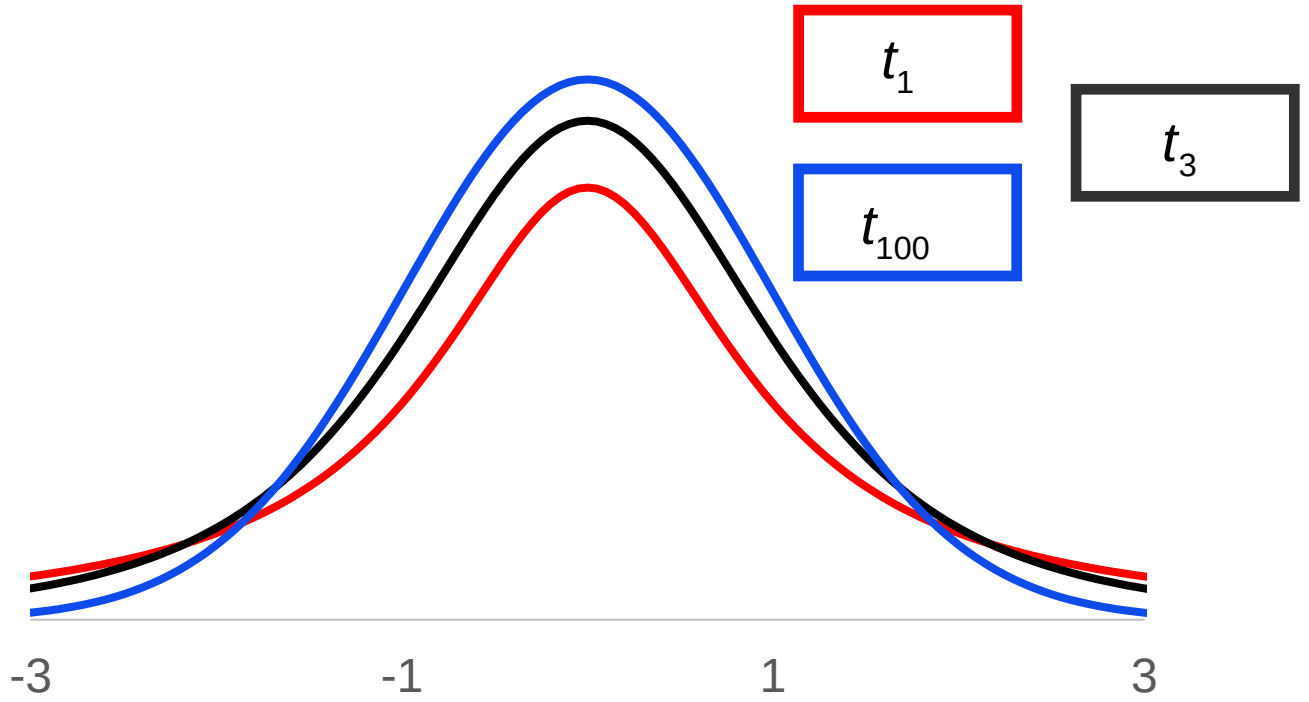
William Sealy Gosset
"Student"
(1876 – 1957)

t-Distribution

- The distribution is symmetric about 0 hence has expectation 0.

- It closely resembles a normal distribution curve, but with slightly heavier tails, making observations far away from 0 slightly more likely than in a normal distribution.

- As the number of degrees of freedom tends to infinity, the distribution converges to that of a standard normal.



t-tests

- Later in this subject, when we revisit the fundamentals of linear regression, we will see the t-distribution.

- If we have a standard regression model $y_i = \alpha + \beta x_i + \varepsilon_i$ where the residuals $\varepsilon_i \sim N(0, \sigma^2)$ with σ^2 unknown, when we estimate this via sample variance, we derive a test statistic for a null hypothesis of a given value for β which is calculated as the ratio of a normal variable to the square root of a chi-squared variable.

- The resulting test is therefore a t-test.

Functions of Normal Variables

- Finally, we consider the ratio of two chi-squared variables, each divided by degrees of freedom.

- Let. $S \sim \chi^2(n)$ and $T \sim \chi^2(m)$ be independent variables. Their joint density is therefore

$$f_{S,T}(s,t) = \begin{cases} \dfrac{1}{\sqrt{2^n}\Gamma\left(\dfrac{n}{2}\right)} e^{-\frac{s}{2}} s^{\frac{n}{2}-1} \dfrac{1}{\sqrt{2^m}\Gamma\left(\dfrac{m}{2}\right)} e^{-\frac{t}{2}} t^{\frac{m}{2}-1} & (s,t) \in [0,\infty)^2 \\ 0 & \text{otherwise} \end{cases}$$

- We now consider the change of variables $X = S$ and $Y = \dfrac{\left(\dfrac{T}{m}\right)}{\left(\dfrac{S}{n}\right)}$ hence $S = X$ and $T = \dfrac{mXY}{n}$.

Functions of Normal Variables

- $S = X$ and $T = \frac{mXY}{n}$ hence $\boldsymbol{J} = \begin{pmatrix} \frac{\partial S}{\partial X} & \frac{\partial S}{\partial Y} \\ \frac{\partial T}{\partial X} & \frac{\partial T}{\partial Y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{mY}{n} & \frac{mX}{n} \end{pmatrix}$ hence $\left|\det(\boldsymbol{J})\right| = \frac{mX}{n}$.

- The joint density of X and Y is then

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\sqrt{2^n}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \frac{1}{\sqrt{2^m}\Gamma\left(\frac{m}{2}\right)} e^{-\frac{mxy}{2n}} \left(\frac{mxy}{n}\right)^{\frac{m}{2}-1} \frac{mx}{n} & (x,y) \in [0,\infty)^2 \\ 0 & \text{otherwise} \end{cases}$$

Functions of Normal Variables

- $$f_{X,Y}(x,y) = \begin{cases} \dfrac{y^{\frac{m}{2}-1}}{\sqrt{2^{m+n}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} e^{-\frac{x}{2}\left(1+\frac{my}{n}\right)} x^{\frac{m+n}{2}-1} & (x,y) \in [0,\infty)^2 \\ 0 & \text{otherwise} \end{cases}$$

- We now integrate to obtain the marginal density of *Y.*

$$f_Y(y) = \begin{cases} \dfrac{y^{\frac{m}{2}-1}\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\sqrt{2^{m+n}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \displaystyle\int_0^\infty e^{-\frac{x}{2}\left(1+\frac{my}{n}\right)} x^{\frac{m+n}{2}-1} dx & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Functions of Normal Variables

- With change of variable $q = \frac{x}{2}\left(1+\frac{my}{n}\right)$ and hence $\frac{dq}{dx} = \frac{1}{2}\left(1+\frac{my}{n}\right)$ we obtain

$$f_Y(y) = \begin{cases} \dfrac{y^{\frac{m}{2}-1}\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\sqrt{2^{m+n}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(\frac{1}{2}\right)^{\frac{m+n}{2}}\left(1+\frac{my}{n}\right)^{\frac{m+n}{2}}}\int_0^\infty e^{-q}q^{\frac{m+n}{2}-1}dq & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Functions of Normal Variables

- Recognising the integral as evaluating to a gamma function, we see that

$$f_Y(y) = \begin{cases} \dfrac{y^{\frac{m}{2}-1}\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\sqrt{2^{m+n}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(\frac{1}{2}\right)^{\frac{m+n}{2}}\left(1+\frac{my}{n}\right)^{\frac{m+n}{2}}} \int_0^\infty e^{-q} q^{\frac{m+n}{2}-1} dq & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

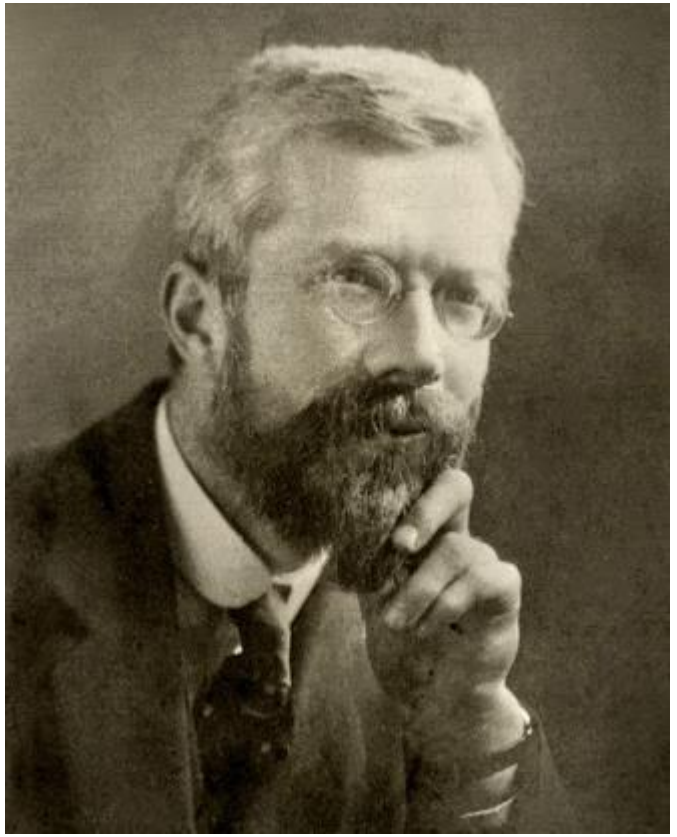
$$\text{gives } f_Y(y) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right) y^{\frac{m}{2}-1}\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{my}{n}\right)^{\frac{m+n}{2}}} & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

F-Distribution

- A variable $Y \sim F(m,n)$ has a **F-distribution** with ***m* and *n* degrees of freedom** and probability density function

$$f_Y(y) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right) y^{\frac{m}{2}-1} \left(\dfrac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)\left(1+\dfrac{my}{n}\right)^{\frac{m+n}{2}}} & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

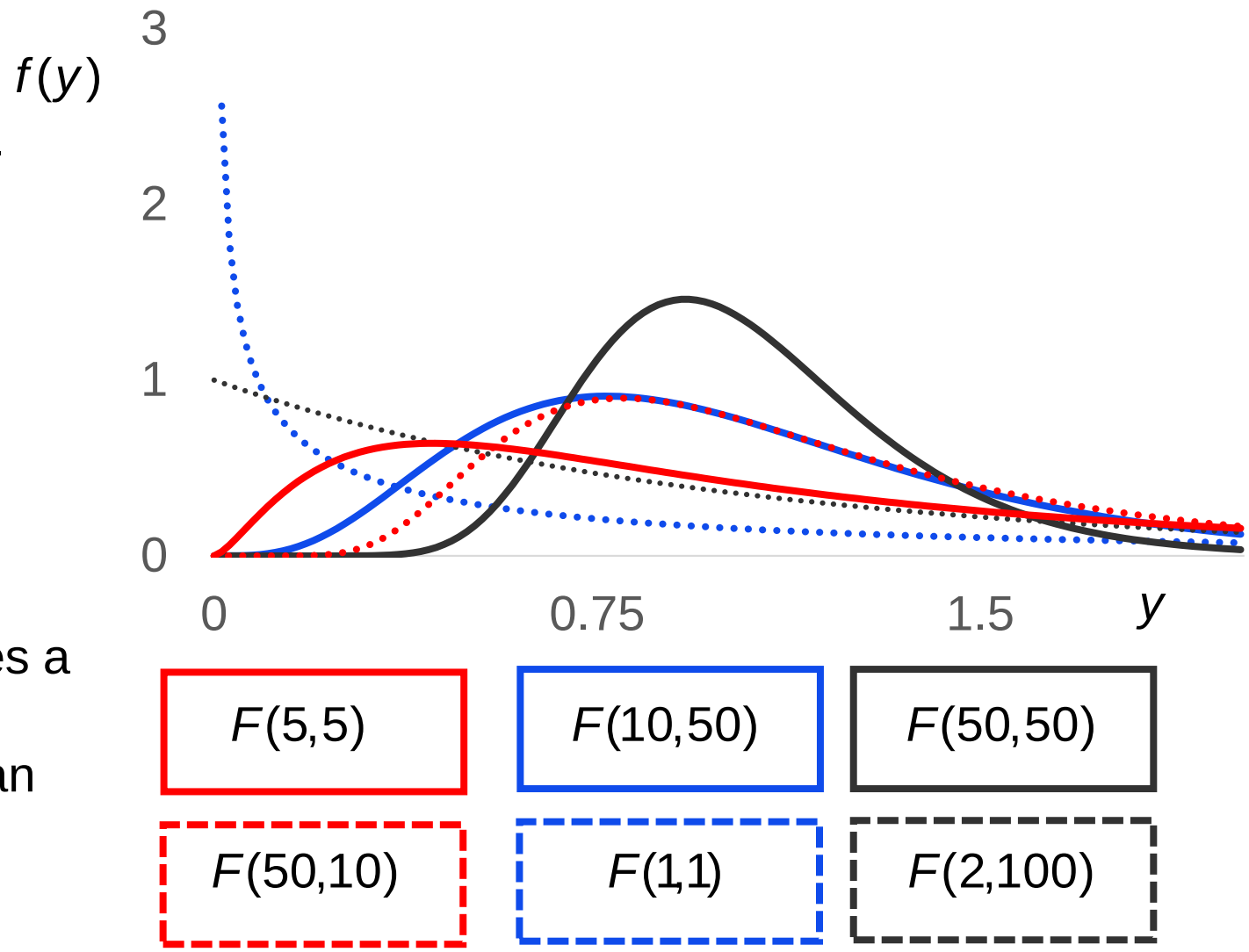
- This is sometimes called a Fisher distribution, after the statistician Ronald Fisher.



Ronald Fisher
(*1890 - 1962*)

F-Distribution

- As it is the ratio of two non-negatively-valued variables, the F-distribution is only non-negatively valued.
- When $m \leq 2$, its mode is at zero. For larger *m*, the mode is positive.
- For large *m* and *n*, the curve resembles a “bell curve” centred around 1. Note, though, that it is not symmetric as it can only take non-negative values.



F-Distribution

- The F-distribution is probably best known for its use in ANOVA (analysis of variance), a technique developed by Ronald Fisher.

- Because it is used to compare the ratio of the variance between treatments to the variance within the same treatment, is it implemented as the ratio of two estimated variances, each of which is chi-squared distributed (assuming normality of residuals in the underlying linear model.)

- We have shown that the ratio of two chi-squared variables is an F-distributed variable.