Comparing Groups: Statistical Tests

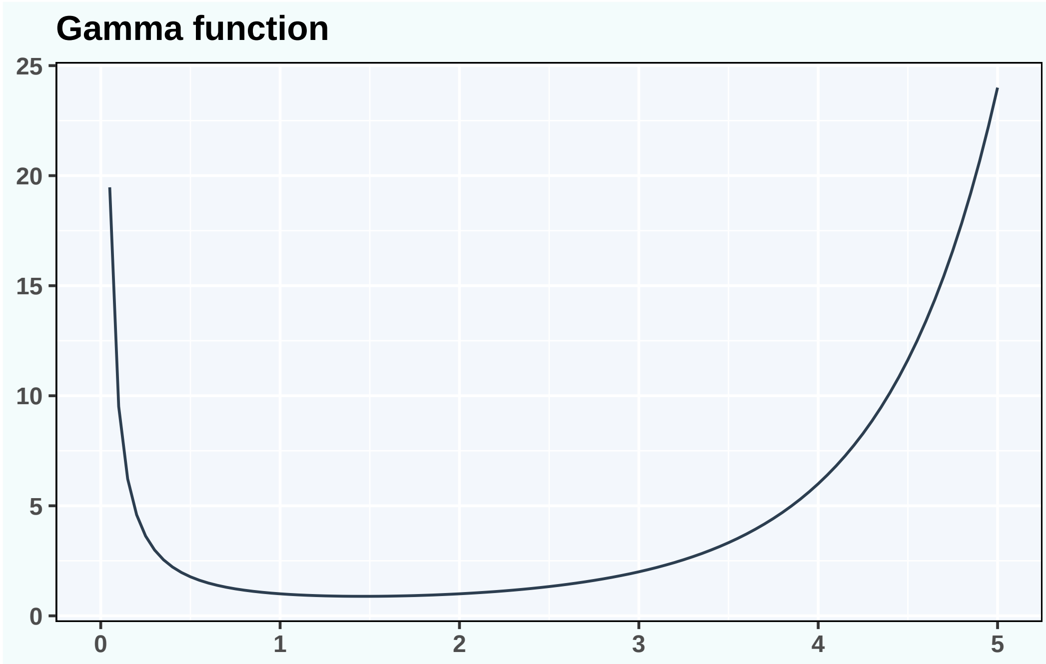
Table of contents

1 Preliminaries

1.1 Gamma function

$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ with $x > 0$[1] and $t > 0$

[1] $\Gamma(x)$ can be defined for $x < 0$ with $x \notin \mathbb{Z}^-$ but we are not interested in those cases



1.1.1 Properties Gamma function

$$\begin{aligned}
\Gamma(\frac{1}{2}) &= \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt \\
&= \int_0^\infty \frac{\sqrt{2}}{z} e^{-\frac{z^2}{2}} z dz \text{ with } t = \frac{z^2}{2} \\
&= \int_0^\infty \sqrt{2}\sqrt{2}\sqrt{\pi} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= 2\sqrt{\pi} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= 2\sqrt{\pi}\frac{1}{2} \\
&= \sqrt{\pi}
\end{aligned}$$

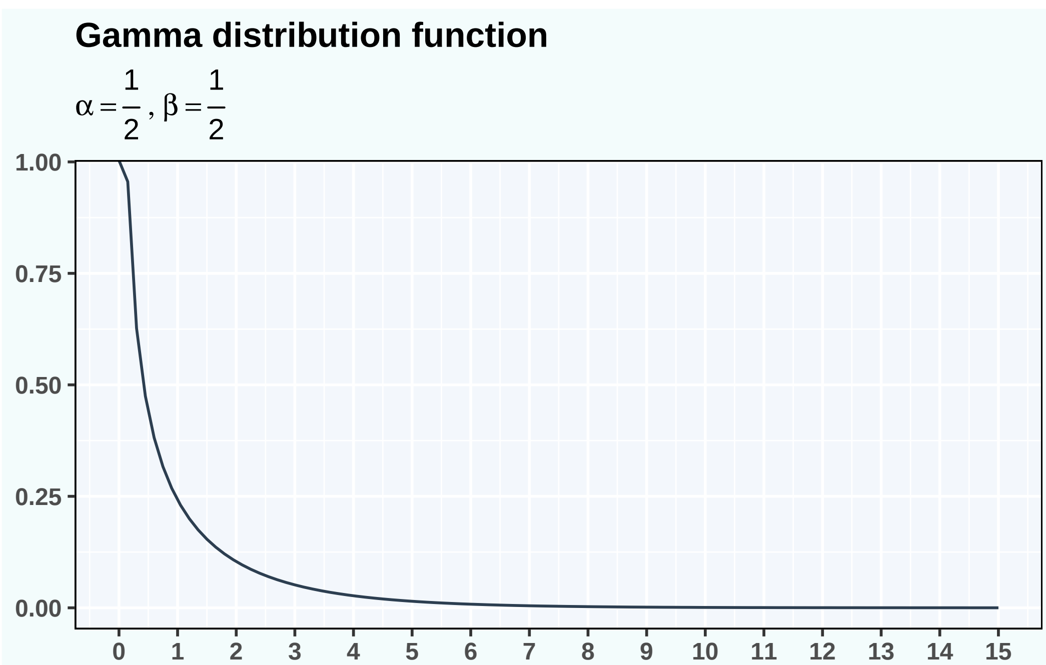
$$
\begin{aligned}
\Gamma(x+1) &= \int_0^\infty t^x e^{-t} dt \\
&= [-t^x e^{-t}]_{t=0}^{t=\infty} + x \int_0^\infty t^{x-1} e^{-t} dt \text{ where we apply integration by parts} \\
&= - \lim_{x\to\infty} \frac{t^x}{e^t} + x\Gamma(x) \\
&= x\Gamma(x) \text{ where we apply L'Hôpital's rule several times}
\end{aligned}
$$

In the case of $x \in \mathbb{N}^*$ we can show that $\Gamma(x+1) = x!$

- For $x = 2$ we have that $\Gamma(2) = 2\Gamma(1) = 2\int_0^\infty e^{-t} = 2[-e^{-x}]_{x=0}^{x=\infty} = 2 \cdot 1$
- Assume $\Gamma(x) = (x-1)!$
- Let $\Gamma(x+1) = x\Gamma(x) = x(x-1)! = x!$

1.2 Gamma distribution function

$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ with $x > 0$, $\alpha > 0$ and $\beta > 0$



Let $f(x; \frac{1}{2}, \frac{1}{2}) = \frac{\frac{1}{2}^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} x^{\frac{1}{2}-1} e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$. Suppose that $Z \sim \mathcal{N}(0,1)$ then it is possible to show that $Z^2 \sim f(x; \frac{1}{2}, \frac{1}{2})$

If F_{X^2} is the cumulative distribution function of X^2 then for any $x \leq 0$ we have that $F_{X^2} = \mathbf{P}(X^2 \leq x) = 0$.

In the case of $x > 0$ we have that $F_{X^2} = \mathbf{P}(X^2 \leq x) = \mathbf{P}(-\sqrt{x} \leq X \leq \sqrt{x}) = 2\mathbf{P}(X < \sqrt{x}) = 2F_X(\sqrt{x})$

We can recover f_{X^2} taking into account that $f_{X^2}(x) = \frac{dF_{X^2}(x)}{dx}$. For $x \leq 0$ we have that $f_{X^2} = 0$ and for $x > 0$ we have that $\frac{dF_{X^2}(x)}{dx} = \frac{d2F_X(\sqrt{x})}{dx} = 2\frac{dF_X(\sqrt{x})}{dx} = 2f_X(\sqrt{x})\frac{1}{2\sqrt{x}} = f_X(\sqrt{x})x^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}} x^{-\frac{1}{2}} = f_{X^2}(x)$

1.2.1 Definition of a chi-squared distribution function

A random variable X follows a chi-squared distribution with k degrees of freedom if it its distribution function is:

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

Where $k > 0$ and $k \in \mathbb{N}$. We will use the notation $X \sim \chi^2(k)$

1.2.2 Moment generating function

The moment generating function of a random variable X is $M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx$

In the case of $X \sim \chi^2(k)$ we have that:

$$\begin{aligned} M_X(t) &= \int_0^{\infty} e^{tx} \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} dx \\ &= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^{\infty} x^{\frac{k}{2}-1} e^{-x(\frac{1}{2}-t)} dx \end{aligned}$$

If $\frac{1}{2} - t < 0$ then $e^{-x(\frac{1}{2}-t)} \longrightarrow \infty$ so the integral diverges in this case and the expectation fails to exist.

If $\frac{1}{2} - t = 0$ then $\int_0^{\infty} x^{\frac{k}{2}-1} dx$ but this means that $\int_0^1 x^{\frac{k}{2}-1} dx + \int_1^{\infty} x^{\frac{k}{2}-1} dx$ don't exist because $\int_1^{\infty} x^{\frac{k}{2}-1} dx$ don't exist taking into account that $\frac{k}{2} - 1 \geq -\frac{1}{2}$

If $\frac{1}{2} - t > 0$ then $\frac{1}{2} > t$. We can define $u = \frac{1}{2} - t$ so

$$
\begin{aligned}
M_X(t) &= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^\infty x^{\frac{k}{2}-1}e^{-x(\frac{1}{2}-t)}dx \\
&= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^\infty \left(\frac{u}{\frac{1}{2}-t}\right)^{\frac{k}{2}-1}\frac{1}{\frac{1}{2}-t}e^{-u}du \text{ with } u = x(\frac{1}{2}-t) \\
&= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})(\frac{1}{2}-t)^{\frac{k}{2}}}\int_0^\infty u^{\frac{k}{2}-1}e^{-u}du \\
&= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})(\frac{1}{2}-t)^{\frac{k}{2}}}\Gamma(\frac{k}{2}) \\
&= \frac{1}{2^{\frac{k}{2}}\frac{(1-2t)^{\frac{k}{2}}}{2^{\frac{k}{2}}}} \\
&= \frac{1}{(1-2t)^{\frac{k}{2}}}
\end{aligned}
$$

1.3 Expected value of a continuous random variable

1.3.1 Joint density functions

Two random variables X and Y are jointly continuous if there is a function $f_{X,Y}(x,y)$ on $\mathbb{R}^2$ such that

$$\mathbf{P}(X \leq s, Y \leq t) = \int_{-\infty}^{t}\int_{-\infty}^{s} f_{X,Y}(x,y)dxdy$$

Such that $f_{X,Y}(x,y) \geq 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$

1.3.2 Marginal densities

If X and Y are jointly continuous with joint density $f_{X,Y}(x,y)$, then the marginal densities are given by

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\
f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx
\end{aligned}
$$

1.3.3 Definition of independence of random continuous variables

Let X, Y be jointly continuous random variables with joint density $f_{X,Y}(x, y)$ and marginal densities $f_X(x)$, $f_Y(y)$. We say they are independent if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

1.3.4 Definition of expected values

If X is a continuous random variable with a probability density function $f(x)$ then

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

If $Z = XY$ where X and Y are continuous independent random variables we have the following result[2]:

$$\begin{aligned}
E[Z] &= \int_{-\infty}^{\infty} zf_Z(z)dz \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x, y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \text{ By independence} \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \text{ By Fubini-Tonelli theorem} \\
&= E[X]E[Y]
\end{aligned}$$

If $Y = g(X)$ where X is a continuous random variable with probability density function $f_X(x)$ there is a theorem which specifies that (Rice 2021, chap. 24, page 122)

$$\begin{aligned}
E[Y] &= \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= \int_{-\infty}^{\infty} g(x)f_X(x)dx
\end{aligned}$$

[2]This proof may have problems and it is not totally rigurous. For example check out the discrete case in https://math.stackexchange.com/questions/3091892/proof-of-exy-ex-ey and the continuous case in https://math.stackexchange.com/questions/3707726/proof-of-expected-value-property-for-product-of-independent-variables

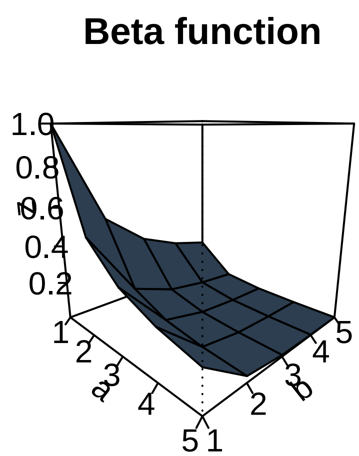
What happen with $W = g(X)h(Y)$ where X and Y are continuous independent random variables with probability density functions $f_X(x)$ and $f_Y(y)$?

We need to establish if $g(X)$ and $h(Y)$ are also independent random variables. If this is the case we have that $E[W] = E[g(X)]E[h(Y)]$.

Also $E[g(X)]E[h(Y)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx \int_{-\infty}^{\infty} h(y) f_Y(y)dy$

1.4 Beta function

$\beta(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ where $a > 0$ and $b > 0$



We have that:

$$\begin{aligned}
\beta(a,b) &= \int_0^1 t^{a-1}(1-t)^{b-1}dt \\
&= \int_1^0 (1-u)^{a-1}u^{b-1} - du \text{ with } u = 1-t \\
&= -\int_1^0 (1-u)^{a-1}u^{b-1}du \\
&= \int_0^1 (1-u)^{a-1}u^{b-1}du \text{ because } \int_b^a f(x) = -\int_a^b f(x) \\
&= \beta(b,a)
\end{aligned}$$

Also because $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$, $\Gamma(b) = \int_0^\infty t^{b-1}e^{-t}dt$ and $\Gamma(a+b) = \int_0^\infty t^{a+b-1}e^{-t}dt$ we have that:

$$\begin{aligned}
\Gamma(a+b)\beta(a,b) &= \Gamma(a+b)\int_0^1 t^{a-1}(1-t)^{b-1}dt \\
&= \Gamma(a+b)\int_0^\infty \frac{u^{a-1}}{(1+u)^{a-1}}\frac{1}{(1-u)^{b-1}}\frac{1}{(1+u)^2}du \text{ where } t = \frac{u}{1+u} \\
&= \Gamma(a+b)\int_0^\infty \frac{u^{a-1}}{(1+u)^{a+b}}du \\
&= \int_0^\infty v^{a+b-1}e^{-v}dv \int_0^\infty \frac{u^{a-1}}{(1+u)^{a+b}}du \\
&= \int_0^\infty u^{a-1}\left(\int_0^\infty \left[\frac{v}{1+u}\right]^{a+b-1}\frac{1}{1+u}e^{-v}dv\right)du \\
&= \int_0^\infty u^{a-1}\left(\int_0^\infty s^{a+b-1}e^{-s(1+u)}ds\right)du \text{ where } s = \frac{v}{1+u} \\
&= \int_0^\infty s^b e^{-s}\left(\int_0^\infty (us)^{a-1}e^{-us}du\right)ds \\
&= \int_0^\infty s^{b-1}e^{-s}\left(\int_0^\infty (us)^{a-1}e^{-us}dus\right)ds \text{ where } dus = sdu \\
&= \Gamma(a)\int_0^\infty s^{b-1}e^{-s}ds \\
&= \Gamma(a)\Gamma(b) \\
\beta(a,b) &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\end{aligned}$$

1.5 Incomplete beta function

$\beta(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ where $x \in [0, 1]$, $a > 0$ and $b > 0$

1.6 Regularized incomplete beta function

$I_x(a, b) = \frac{\beta(x;a,b)}{\beta(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}dt$ where $x \in [0, 1]$, $a > 0$ and $b > 0$

$$\begin{aligned}
1 - I_{1-x}(b, a) &= \frac{\beta(1-x; b, a)}{\beta(b, a)} \\
&= 1 - \frac{1}{\beta(b, a)} \int_0^{1-x} t^{b-1}(1-t)^{a-1}dt \\
&= 1 - \frac{1}{\beta(b, a)} \int_1^x (1-u)^{b-1}u^{a-1}(-du) \text{ where } u = 1 - t \\
&= \frac{\beta(a, b)}{\beta(a, b)} + \frac{1}{\beta(a, b)} \int_1^x u^{a-1}(1-u)^{b-1}dt \\
&= \frac{1}{\beta(a, b)} \int_0^1 u^{a-1}(1-u)^{b-1}dt + \frac{1}{\beta(a, b)} \int_1^x u^{a-1}(1-u)^{b-1}dt \\
&= \frac{1}{\beta(a, b)} \int_0^x u^{a-1}(1-u)^{b-1}dt \\
&= \frac{\beta(x; a, b)}{\beta(a, b)} \\
&= I_x(a, b)
\end{aligned}$$

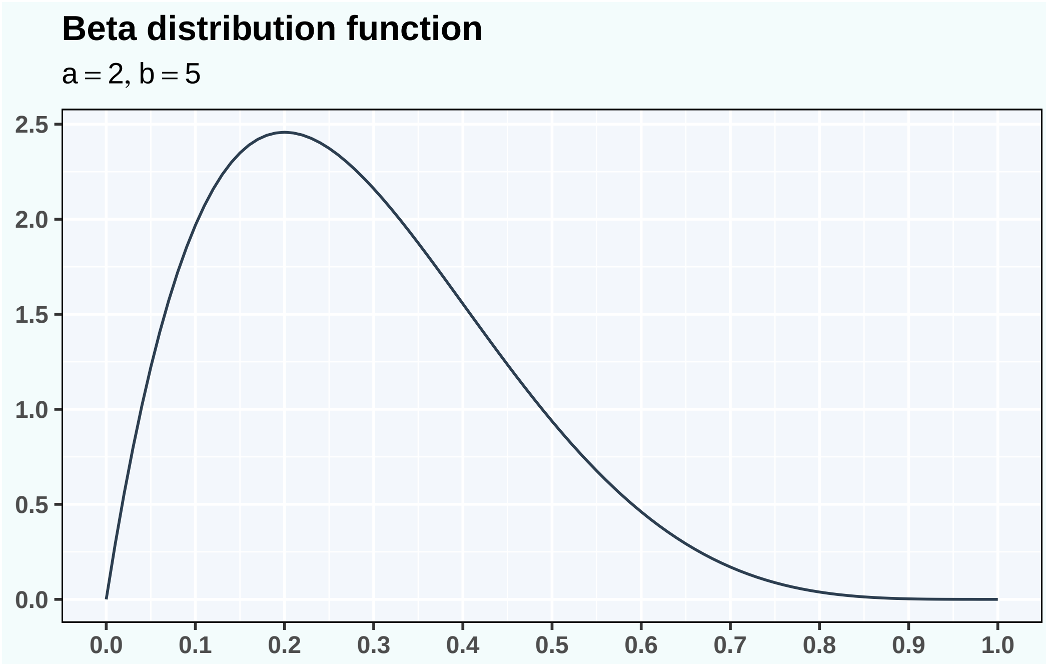
If $p \in [0, 1]$, $n \geq x \geq 0$, $n \in \mathbb{N}^*$ and $x \in \mathbb{N}^*$ we can define:

$$
\begin{aligned}
I_p(x, n-x+1) &= \frac{\beta(p; x, n-x+1)}{\beta(x, n-x+1)} \\
&= \frac{\Gamma(x+n-x+1)}{\Gamma(x)\Gamma(n-x+1)} \int_0^p t^{x-1}(1-t)^{n-x+1-1} dt \\
&= \frac{n!}{(x-1)!(n-x)!} \int_0^p t^{x-1}(1-t)^{n-x} dt \\
&= x\frac{n!}{x!(n-x)!} \int_0^p t^{x-1}(1-t)^{n-x} dt \\
&= x\binom{n}{x} \int_0^p t^{x-1}(1-t)^{n-x} dt \\
\frac{dI_p(x, n-x+1)}{dp} &= x\binom{n}{x} \frac{d}{dp} \int_0^p t^{x-1}(1-t)^{n-x} dt \\
&= x\binom{n}{x} p^{x-1}(1-p)^{n-x} \text{ By the fundamental theorem of calculus}
\end{aligned}
$$

$$
\begin{aligned}
I_{1-p}(n-x, x+1) &= 1 - I_p(x+1, n-x) \\
&= 1 - \frac{\beta(p; x+1, n-x)}{\beta(x+1, n-x)} \\
&= 1 - \frac{\Gamma(x+1+n-x)}{\Gamma(x+1)\Gamma(n-x)} \int_0^p t^x(1-t)^{n-x-1} dt \\
&= 1 - \frac{n!}{x!(n-x-1)!} \int_0^p t^x(1-t)^{n-x-1} dt \\
&= 1 - (n-x)\frac{n!}{x!(n-x)!} \int_0^p t^x(1-t)^{n-x-1} dt \\
&= 1 - (n-x)\binom{n}{x} \int_0^p t^x(1-t)^{n-x-1} dt \\
\frac{dI_{1-p}(n-x, x+1)}{dp} &= -(n-x)\binom{n}{x} \frac{d}{dp} \int_0^p t^x(1-t)^{n-x-1} dt \text{ By the fundamental theorem of calculus} \\
&= -(n-x)\binom{n}{x} p^x(1-p)^{n-x-1}
\end{aligned}
$$

1.7 Beta distribution

$f(x; a, b) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ where $x \in [0,1]$, $a > 0$ and $b > 0$



We use the notation $X \sim Beta(a, b)$ to say that the random variable X follows a beta distribution with parameters $a > 0$ and $b > 0$

1.7.1 Cumulative distribution function

$$\begin{aligned}
F(x; a, b) &= \mathbf{P}(X \leq x) \\
&= \int_{-\infty}^{x} \frac{1}{\beta(a, b)} t^{a-1}(1-t)^{b-1} dt \\
&= \frac{1}{\beta(a, b)} \left[\int_{-\infty}^{0} t^{a-1}(1-t)^{b-1} dt + \int_{0}^{x} t^{a-1}(1-t)^{b-1} dt \right] \\
&= \frac{1}{\beta(a, b)} \int_{0}^{x} t^{a-1}(1-t)^{b-1} dt \\
&= \frac{\beta(x, a, b)}{\beta(a, b)} \\
&= I_x(a, b)
\end{aligned}$$

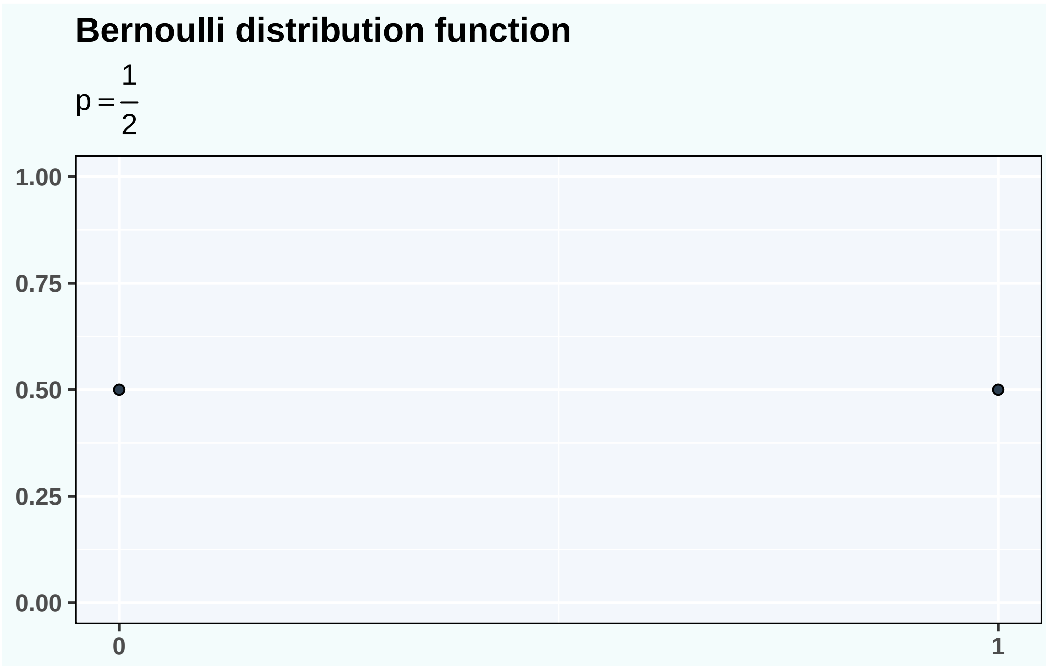
1.8 Binomial distribution

A random variable X follows a binomial distribution if its distribution function is:

$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$

Where $0 < p < 1$, $n \in \{0, 1, ...\}$, $x \leq n$ and $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

We use the notation $X \sim B(n, p)$



If $n = 1$ we say that X follows a bernoulli distribution:

$f(x; p) = p^x (1-p)^{1-x}$

Where $x \in 0, 1$

We use the notation $X \sim Bernoulli(p)$

1.8.1 Binomial coefficients

$i\binom{n}{i} = i\frac{n!}{i!(n-i)!} = i\frac{n(n-1)!}{i(i-1)!(n-1-(i-1))!} = n\frac{(n-1)!}{(i-1)!((n-1)-(i-1))!} = n\binom{n-1}{i-1}$

$(n-i)\binom{n}{i} = (n-i)\frac{n!}{i!(n-i)!} = (n-i)\frac{n(n-1)!}{i!(n-i)(n-i-1)!} = n\frac{(n-1)!}{i!((n-1)-i)!} = n\binom{n-1}{i}$

1.8.2 Cumulative distribution function

$$
\begin{aligned}
F(x;n,p) &= \mathbf{P}(X \le x) \\
&= \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p^1 (1-p)^{n-1} + \cdots + \binom{n}{x-1} p^{x-1} (1-p)^{n-(x-1)} + \binom{n}{x} p^x (1-p)^{n-x} \\
&= \binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p^1 (1-p)^{n-1} + \cdots + \binom{n}{x-1} p^{x-1} (1-p)^{n-(x-1)} + \binom{n}{x} p^x (1-p)^{n-x} \\
\frac{dF(x;n,p)}{dp} &= \binom{n}{0} (n-0) p^0 (1-p)^{n-1} + \binom{n}{1} (1p^{1-1}(1-p)^{n-1} - (n-1)p^1(1-p)^{n-2}) + \cdots + \\
&\binom{n}{x-1} ((x-1)p^{x-2}(1-p)^{n-(x-1)} - (n-(x-1))p^{x-1}p^{n-(x-1)}) + \\
&\binom{n}{x} (xp^{x-1}(1-p)^{n-x} - (n-x)p^x(1-p)^{n-x-1}) \\
&= \sum_{i=1}^{x} \binom{n}{i} i p^{i-1} (1-p)^{n-i} - \sum_{i=0}^{x} \binom{n}{i} (n-i) p^i (1-p)^{n-i-1} \\
&= n \left[\sum_{i=1}^{x} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} - \sum_{i=0}^{x} \binom{n-1}{i} p^i (1-p)^{n-i-1} \right] \\
&= n \left[\sum_{i=1}^{x} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} - \sum_{i=1}^{x} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} - \binom{n-1}{x} p^x (1-p)^{n-x-1} \right] \\
&= -n \binom{n-1}{x} p^x (1-p)^{n-x-1} \\
&= -n \frac{(n-1)!}{x!(n-1-x)!} p^x (1-p)^{n-x-1} \\
&= -(n-x) \binom{n}{x} p^x (1-p)^{n-x-1}
\end{aligned}
$$

Taking into account that $\frac{dF(x;n,p)}{dp} = \frac{dI_{1-p}(n-x,x+1)}{dp}$

1.8.2.1 Relation with the Beta cumulative distribution function

- According to (Johnson, Kemp, and Kotz 2005, 119) we have the following result:

$$
\begin{aligned}
Pr[X \geq x] &= \sum_{i=x}^{n} \binom{n}{i} p^i (1-p_i)^{n-i} \\
&= I_p(x, n-x+1) \\
&= \frac{\beta(p; x, n-x+1)}{\beta(x, n-x+1)} \\
&= \frac{\int_0^p t^{x-1}(1-t)^{n-x} dt}{\int_0^1 t^{x-1}(1-t)^{n-x} dt}
\end{aligned}
$$

- Using this result we have also the following result:

$$
\begin{aligned}
Pr[X \leq x] &= 1 - Pr[X \geq x+1] \\
&= 1 - I_p(x+1, n-(x+1)+1) \\
&= 1 - I_p(x+1, n-x) \\
&= 1 - \frac{\beta(p; x+1, n-x)}{\beta(x+1, n-x)} \\
&= 1 - \frac{\int_0^p t^{x}(1-t)^{n-x-1} dt}{\int_0^1 t^{x}(1-t)^{n-x-1} dt}
\end{aligned}
$$

Let's see the application of this result for $Pr[X \geq x]$:

```
pbinom(q = 156, size = 300, prob = 0.5,
       lower.tail = FALSE)
```

```
[1] 0.2264879
```

```
pbeta(q = 0.5, shape1 = 157, shape2 = 300 - 157 + 1,
      lower.tail = TRUE)
```

```
[1] 0.2264879
```

The first part in relation to $q = 156$ it is fixed in that way taking into account that `lower.tail = FALSE` means that $Pr[X > x]$. Therefore $Pr[X > 156] = Pr[X \geq 157]$

Let's see the application of this result for $Pr[X \leq x]$:

```r
pbinom(q = 157, size = 300, prob = 0.5,
       lower.tail = TRUE)
```

```
[1] 0.8067451
```

```r
1 - pbeta(q = 0.5, shape1 = 157 + 1, shape2 = 300 - 157,
          lower.tail = TRUE)
```

```
[1] 0.8067451
```

1.8.3 Moment generating function

$$
\begin{aligned}
M_X(t) &= E[e^{tX}] \\
&= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{(n-x)} \\
&= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{(n-x)} \\
&= (pe^t + 1 - p)^n \text{ By the binomial theorem}
\end{aligned}
$$

If $X_1, \ldots X_n$ are independent random variables with $X_i \sim Bernoulli(p)$ then:

$$\sum_{i=1}^{n} X_i \sim B(n, p)$$

We can proof this with the moment generating function

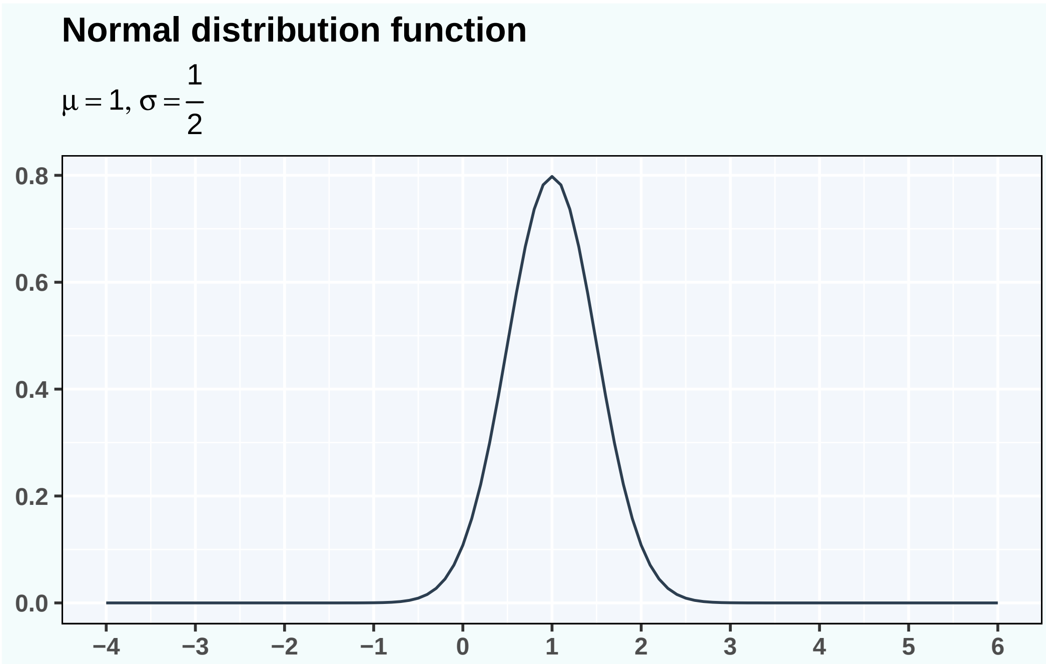
$$
\begin{aligned}
M_{\sum_{i=1}^{n} X_i}(t) &= E[e^{t(\sum_{i=1}^{n} X_i)}] \\
&= E[e^{tX_1} \ldots e^{tX_n)}] \\
&= E[e^{tX_1}] \cdots E[e^{tX_n}] \text{ Because } X_1, \ldots X_k \text{ are independent random variables} \\
&= (pe^t + 1 - p) \ldots (pe^t + 1 - p) \text{ Because } X_i \sim Bernoulli(p) \\
&= (pe^t + 1 - p)^n
\end{aligned}
$$

1.9 Normal distribution

A random variable X follows a normal distribution if its distribution function is:

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

We say that $X \sim \mathcal{N}(\mu, \sigma)$



1.9.1 Cumulative distribution function

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} \sqrt{2}\sigma dt \text{ where } t = \frac{x-\mu}{\sqrt{2}\sigma} \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} dt \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} dt \\
&= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} e^{-u^2} du + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} dt \text{ where } t = -u \\
&= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} e^{-u^2} du + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} dt \text{ where } erf(y) = \frac{2}{\sqrt{\pi}} \int_{0}^{y} e^{-t^2} dt \\
&= \frac{1}{2} \lim_{y\to\infty} erf(y) + \frac{1}{2} erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\
&= \frac{1}{2} + \frac{1}{2} erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \text{ By the properties of } erf \\
&= \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]
\end{aligned}
$$

1.10 Student's t-distribution

Let $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2(\nu)$ and Z and V independent random variables then $Y \sim \frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t(\nu)$ where t is the t-distribution with $\nu > 0$ degrees of freedom

So we have that $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ and $f(v) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}}$

Let $f(z,v)$ by the joint probability density function of Z and V. Because Z and V are independent random variables we have that:

$$
\begin{aligned}
f(z,v) &= f(z)f(v) \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}}
\end{aligned}
$$

We can specify a variable $t = \frac{z}{\sqrt{\frac{v}{\nu}}}$ such that $z = t\sqrt{\frac{v}{\nu}}$ and $dz = \sqrt{\frac{v}{\nu}}dt$. So:

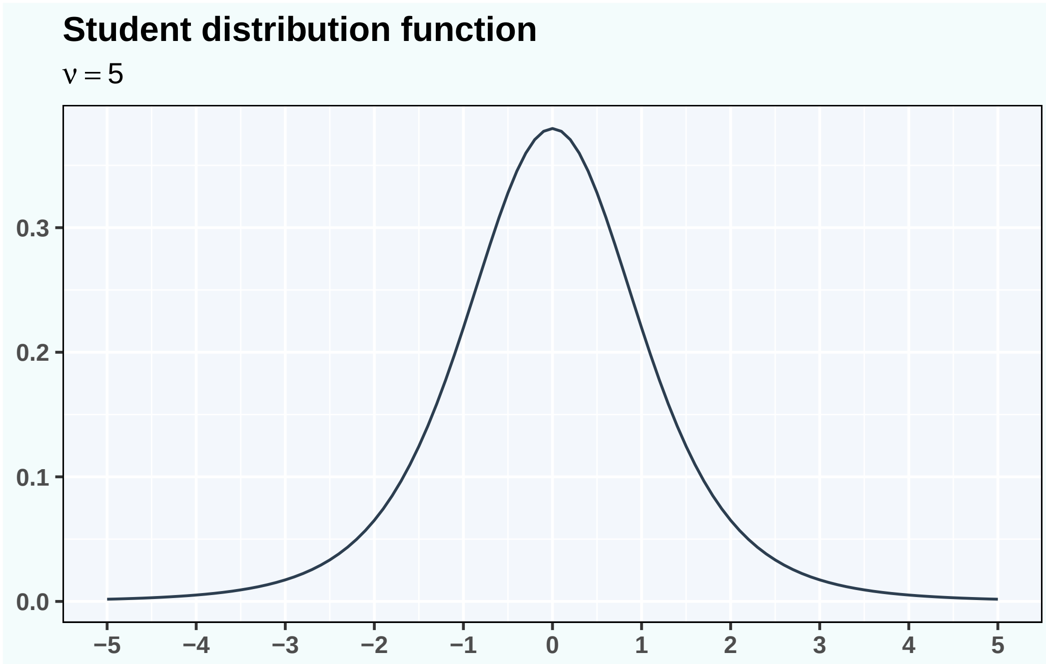
$$\begin{aligned}
F_Z(z) &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2} dw \\
F_Z(t) &= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2\frac{v}{\nu}} \sqrt{\frac{v}{\nu}} ds \\
\frac{F_Z(t)}{dt} &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2\frac{v}{\nu}} \sqrt{\frac{v}{\nu}} \\
f_t(t) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2\frac{v}{\nu}} \sqrt{\frac{v}{\nu}}
\end{aligned}$$

Also we have that $f(z,v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}}$ for $v > 0$ and $z \in (-\infty, \infty)$. In another case $f(z,v) = 0$. Therefore we have that:

$$\begin{aligned}
f(z,v) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}} \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2\frac{v}{\nu}} \sqrt{\frac{v}{\nu}} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}} \\
&= \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{1}{2}t^2\frac{v}{\nu}} v^{\frac{1}{2}} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}} \\
&= \frac{1}{\sqrt{2\pi\nu} 2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu-1}{2}} e^{-\frac{v}{2}(1+\frac{t^2}{\nu})}
\end{aligned}$$

Where $-\infty < t < \infty$ and $v > 0$. So with this joint probability density function we can recover $f_T(t)$ using the definition of the margina density:

$$
\begin{aligned}
f_T(t) &= \int_0^\infty \frac{1}{\sqrt{2\pi\nu}2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} v^{\frac{\nu-1}{2}} e^{-\frac{v}{2}(1+\frac{t^2}{\nu})} dv \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\nu}2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} 2^{\frac{\nu-1}{2}} w^{\frac{\nu-1}{2}} (1+\frac{t^2}{\nu})^{-\frac{\nu-1}{2}} e^{-w} 2(1+\frac{t^2}{\nu})^{-1} dw \text{ where } w = \frac{v}{2}(1+\frac{t^2}{\nu}) \\
&= \int_0^\infty \frac{1}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} w^{\frac{\nu-1}{2}} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} e^{-w} dw \\
&= \frac{1}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \int_0^\infty w^{\frac{\nu-1}{2}} e^{-w} dw \\
&= \frac{1}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \int_0^\infty w^{\frac{\nu+1}{2}-1} e^{-w} dw \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \\
&= \frac{1}{\sqrt{\nu}\frac{\sqrt{\pi}\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})}} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \\
&= \frac{1}{\sqrt{\nu}\frac{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})}} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \\
&= \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})} (1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}
\end{aligned}
$$



Student distribution function
ν = 5
0.3
0.2
0.1
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5

1.10.1 Cumulative distribution function[3]

$$
\begin{aligned}
F_\nu(t) &= \int_{-\infty}^{t} \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})}(1+\frac{s^2}{\nu})^{-\frac{\nu+1}{2}} ds \\
&= \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})} \int_{-\infty}^{t} (1+\frac{s^2}{\nu})^{-\frac{\nu+1}{2}} ds \\
&= \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})} \int_{-\infty}^{t} (\frac{\nu+s^2}{\nu})^{-\frac{\nu+1}{2}} ds \\
&= \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})} \int_{0}^{\frac{\nu}{\nu+t^2}} (\frac{1}{x})^{-\frac{\nu+1}{2}} \frac{\nu}{2x^2\sqrt{\nu(\frac{1-x}{x})}} dx \text{ where } x = \frac{\nu}{\nu+s^2} \text{ and } s<0 \\
&= \frac{1}{2}\frac{1}{\beta(\frac{1}{2},\frac{\nu}{2})} \int_{0}^{\frac{\nu}{\nu+t^2}} x^{\frac{\nu}{2}-1}(1-x)^{-\frac{1}{2}} dx \\
&= \frac{1}{2}\frac{1}{\beta(\frac{1}{2},\frac{\nu}{2})} \int_{0}^{\frac{\nu}{\nu+t^2}} x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1} dx \\
&= \frac{1}{2}\frac{\beta(\frac{\nu}{\nu+t^2};\frac{\nu}{2},\frac{1}{2})}{\beta(\frac{1}{2},\frac{\nu}{2})} \\
&= \frac{1}{2}\frac{\beta(\frac{\nu}{\nu+t^2};\frac{\nu}{2},\frac{1}{2})}{\beta(\frac{\nu}{2},\frac{1}{2})} \\
&= \frac{1}{2} I_{\frac{\nu}{\nu+t^2}}(\frac{\nu}{2},\frac{1}{2})
\end{aligned}
$$

In the case of $t \geq 0$ check out (Johnson, Kotz, and Balakrishnan 1995, 2:364):

[3]You can check out the derivantion in https://statproofbook.github.io/P/f-pdf

$$
\begin{aligned}
F_\nu(t) &= F_\nu(0) + F_\nu(t) - F_\nu(0) \\
&= \frac{1}{2} + \int_0^t \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})}(1+\frac{s^2}{\nu})^{-\frac{\nu+1}{2}} ds \\
&= \frac{1}{2} + \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})}\int_1^{\frac{\nu}{\nu+t^2}} (\frac{1}{x})^{-\frac{\nu+1}{2}} \frac{-\nu}{2x^2\sqrt{\nu(\frac{1-x}{x})}} dx \text{ where } x = \frac{\nu}{\nu+s^2} \text{ and } s \geq 0 \\
&= \frac{1}{2} - \frac{1}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})}\int_{\frac{\nu}{\nu+t^2}}^1 (\frac{1}{x})^{-\frac{\nu+1}{2}} \frac{-\nu}{2x^2\sqrt{\nu(\frac{1-x}{x})}} dx \\
&= \frac{1}{2} + \frac{1}{2}\frac{1}{\beta(\frac{1}{2},\frac{\nu}{2})}\int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{-\frac{1}{2}} dx \\
&= \frac{1}{2} + \frac{1}{2}\frac{1}{\beta(\frac{\nu}{2},\frac{1}{2})}\int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1} dx
\end{aligned}
$$

We have the following result:

$$
\begin{aligned}
\frac{\int_0^{\frac{\nu}{\nu+t^2}} x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1} + \int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1}}{\beta(\frac{\nu}{2},\frac{1}{2})} &= 1 \\
\frac{\int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1}}{\beta(\frac{\nu}{2},\frac{1}{2})} &= 1 - \frac{\int_0^{\frac{\nu}{\nu+t^2}} x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1}}{\beta(\frac{\nu}{2},\frac{1}{2})} \\
\frac{\int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1}}{\beta(\frac{\nu}{2},\frac{1}{2})} &= 1 - I_{\frac{\nu}{\nu+t^2}}(\frac{\nu}{2},\frac{1}{2})
\end{aligned}
$$

Using the above result we have the following:

$$
\begin{aligned}
F_\nu(t) &= \frac{1}{2} + \frac{1}{2}\frac{1}{\beta(\frac{\nu}{2},\frac{1}{2})}\int_{\frac{\nu}{\nu+t^2}}^1 x^{\frac{\nu}{2}-1}(1-x)^{\frac{1}{2}-1} dx \\
&= \frac{1}{2} + \frac{1}{2}\left[1 - I_{\frac{\nu}{\nu+t^2}}(\frac{\nu}{2},\frac{1}{2})\right] \\
&= 1 - \frac{1}{2} I_{\frac{\nu}{\nu+t^2}}(\frac{\nu}{2},\frac{1}{2}) \text{ where } t \geq 0
\end{aligned}
$$

1.11 Transformation of random variables

1.11.1 Transformations of a single random variable

Assume that X is a random variable where X is real-valued. In that sense $X : \Omega \longrightarrow \mathbb{R}$ where Ω is the sample space. Now let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous function. Then we can define a random variable $Y = g(X)$ where the idea is to find the cumulative distribution function $F_Y(y)$ given the cumulative distribution function $F_X(y)$

The classical example is the case where $X \sim \mathcal{N}(0,1)$ where we can do find $Y = X^2$. We can proceed in the following way:

$$
\begin{aligned}
F_Y(y) &= Pr[Y \leq y] \\
&= Pr[X^2 \leq y] \\
&= Pr[|X| \leq \sqrt{y}] \\
&= Pr[-\sqrt{y} \leq X \leq \sqrt{y}] \\
&= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2} dx \\
&= 2\int_{0}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\
&= 2\int_{0}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u} \frac{1}{2\sqrt{u}} du \text{ where } u = x^2 \text{ with } u > 0 \\
&= \int_{0}^{y} \frac{1}{\sqrt{2u\pi}} e^{-\frac{1}{2}u} du \\
\frac{dF_Y(y)}{dy} &= \int_{0}^{y} \frac{1}{\sqrt{2u\pi}} e^{-\frac{1}{2}u} du \\
f_Y(y) &= \frac{1}{\sqrt{2y\pi}} e^{-\frac{1}{2}y} \\
&= \frac{1}{2^{\frac{1}{2}}\sqrt{\pi}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} \\
&= \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{y}{2}} \text{ where } y > 0 \\
&= \chi_1^2
\end{aligned}
$$

Now assume g is strictly increasing, $x_1 < x_2 \Longrightarrow g(x_1) < g(x_2)$, and differentiable function.

We have that g is a one-to-one function and its inverse g^{-1} is also a function and one-to-one. Furthermore, g^{-1} is also strictly increasing

$$
\begin{aligned}
F_Y(y) &= Pr[Y < y] \\
&= Pr[g(X) < y] \\
&= Pr[X < g^{-1}(y)] \\
&= \int_{-\infty}^{g^{-1}(y)} f_X(x)dx \\
&= \int_{-\infty}^{y} f_X(g^{-1}(t))(g^{-1})'(t)dt \text{ where } x = g^{-1}(t) \\
&= \int_{-\infty}^{y} f_X(g^{-1}(t))\frac{1}{g'(g^{-1}(t))}dt \text{ where } g'(g^{-1}(t)) \neq 0 \\
\frac{dF_Y(y)}{dy} &= f_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))} \\
f_Y(y) &= f_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}
\end{aligned}
$$

We can check the same result for g when is strictly decreasing, $x_1 < x_2 \Longrightarrow g(x_1) > g(x_2)$, and differentiable function.

We have that g is a one-to-one function and its inverse g^{-1} is also a function and one-to-one. Furthermore, g^{-1} is also strictly decreasing

$$
\begin{aligned}
F_Y(y) &= Pr[Y < y] \\
&= Pr[g(X) < y] \\
&= Pr[X > g^{-1}(y)] \\
&= \int_{g^{-1}(y)}^{\infty} f_X(x)dx \\
&= \int_{y}^{\infty} f_X(g^{-1}(t))(g^{-1})'(t)dt \text{ where } x = g^{-1}(t) \\
&= \int_{y}^{\infty} f_X(g^{-1}(t))\frac{1}{g'(g^{-1}(t))}dt \text{ where } g'(g^{-1}(t)) \neq 0 \\
&= 1 - \int_{-\infty}^{y} f_X(g^{-1}(t))\frac{1}{g'(g^{-1}(t))}dt \\
\frac{dF_Y(y)}{dy} &= f_X(g^{-1}(y))\frac{-1}{g'(g^{-1}(y))} \\
f_Y(y) &= f_X(g^{-1}(y))\frac{-1}{g'(g^{-1}(y))}
\end{aligned}
$$

Taking into account both results we have that:

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))|}$$

We can apply this result to the following case where $X \sim \mathcal{N}(\mu, \sigma^2)$ and the idea is to find the distribution of $Y = e^X$

- $\frac{de^x}{dx} = e^x$
- $g^{-1}(y) = log_e y$
- $\frac{de^x}{dx} = e^x \Big|_{x=log_e y} = y > 0$
- $f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{log_e y - \mu)}{\sigma}} \frac{1}{y} = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{log_e y - \mu)}{\sigma}}$

Now we can generalize for X_1 and X_2 to find the probability density function using the concept of **Jacobian**

Let $f_{X_1,X_2}(x_1, x_2)$ the probability density function, $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ then we have:

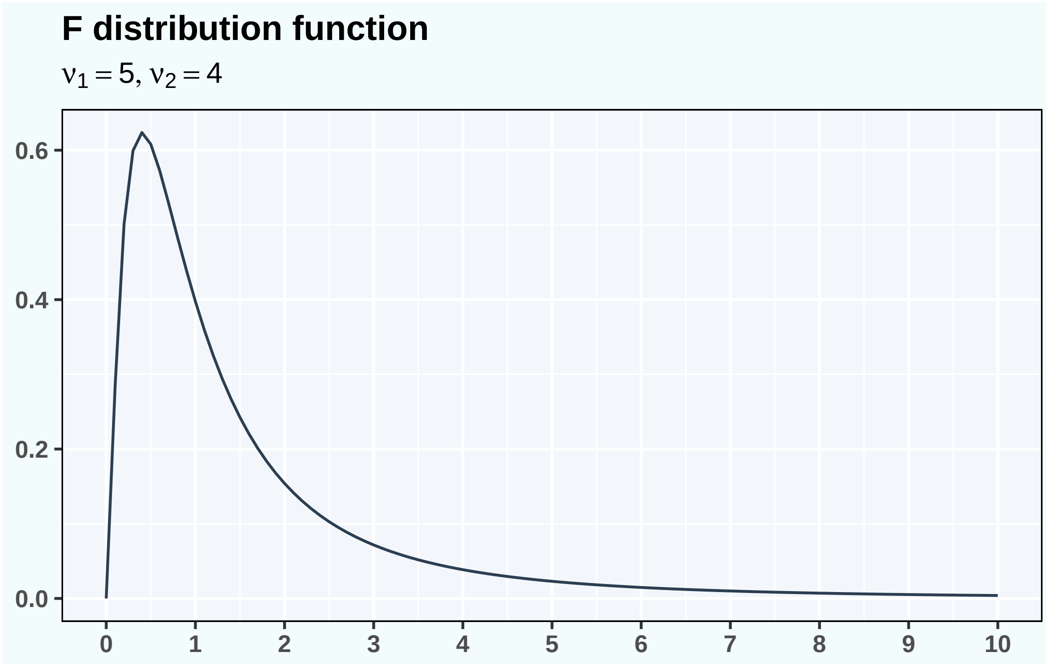
- $y_1 = g_1(x_1, x_2)$, $y_2 = g_2(x_1, x_2)$
- We need to solve for x_1 and x_2 in terms of y_1 and y_2
  - $x_1 = h(y_1, y_2)$
  - $x_2 = h(y_1, y_2)$
- We need to compute the following Jacobin:

$$\mathbf{J} = \begin{bmatrix} \frac{dx_1}{dy_1} & \frac{dx_1}{dy_2} \\ \frac{dx_2}{dy_1} & \frac{dx_2}{dy_2} \end{bmatrix}$$

- Then we need to calculate $|det(\mathbf{J})|$, that is the absolute value of the determinant of the Jacobian, $\left|\frac{dx_1}{dy_1}\frac{dx_2}{dy_2} - \frac{dx_1}{dy_2}\frac{dx_2}{dy_1}\right|$
- Finally we have that $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(h(y_1, y_2), h(y_1, y_2))|det(\mathbf{J})|$

1.12 F-distribution

If X_1 and X_2 are independent random variables distributed as $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ then the distribution of $\frac{\frac{X_1}{\nu_1}}{\frac{X_2}{\nu_2}}$ is the F-distribution with ν_1 and ν_2 degrees of freedom.



1.12.1 Probability density function [4]

- Let $F = \frac{\frac{X_1}{\nu_1}}{\frac{X_2}{\nu_2}}$ and $W = Y$
- We have that $f = \frac{\frac{x_1}{\nu_1}}{\frac{x_2}{\nu_2}}$ and $w = y$. So $x_1 = fw\frac{\nu_1}{\nu_2}$ and $x_2 = w$
- The Jacobian is defined as:

$$\mathbf{J} = \begin{bmatrix} w\frac{\nu_1}{\nu_2} & f\frac{\nu_1}{\nu_2} \\ 0 & 1 \end{bmatrix}$$

- We have that $|det(\mathbf{J})| = |w\frac{\nu_1}{\nu_2}| = w\frac{\nu_1}{\nu_2}$ because w is a realization of a random variable that it is distributed as a $\chi^2_{\nu_2}$, $\nu_1 > 0$ and $\nu_2 > 0$
- The we have that:

[4] Check this to specify the null hypothesis and calculate the expected values https://online.stat.psu.edu/stat500/lesson/8/8.1

$$
\begin{aligned}
f_{F,W}(f,w) &= f_{X_1,X_2}(fw\frac{\nu_1}{\nu_2}, w)|det(\mathbf{J})| \\
&= f_{X_1}(fw\frac{\nu_1}{\nu_2})f_{X_2}(w)|det(\mathbf{J})| \\
&= \frac{1}{2^{\frac{\nu_1}{2}}\Gamma(\frac{\nu_1}{2})}(fw\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}-1}e^{-fw\frac{\nu_1}{2\nu_2}}\frac{1}{2^{\frac{\nu_2}{2}}\Gamma(\frac{\nu_2}{2})}w^{\frac{\nu_2}{2}-1}e^{-\frac{w}{2}}w\frac{\nu_1}{\nu_2} \\
&= \frac{(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}}{2^{\frac{\nu_1+\nu_2}{2}}\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}w^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{w}{2}(f\frac{\nu_1}{\nu_2}+1)}
\end{aligned}
$$

Now we need to recover $f_F(f)$ using the probability marginal density function:

$$
\begin{aligned}
f_F(f) &= \int_0^\infty f_{F,W}(f,w)dw \\
&= \int_0^\infty \frac{(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}}{2^{\frac{\nu_1+\nu_2}{2}}\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}w^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{w}{2}(f\frac{\nu_1}{\nu_2}+1)}dw \\
&= \frac{(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}}{2^{\frac{\nu_1+\nu_2}{2}}\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\int_0^\infty w^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{w}{2}(f\frac{\nu_1}{\nu_2}+1)}dw \\
&= \frac{(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}}{2^{\frac{\nu_1+\nu_2}{2}}\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{(\frac{1}{2}(f\frac{\nu_1}{\nu_2}+1))^{\frac{\nu_1+\nu_2}{2}}}\int_0^\infty \frac{w^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{w}{2}(f\frac{\nu_1}{\nu_2}+1)}(f\frac{\nu_1}{\nu_2}+1))^{\frac{\nu_1+\nu_2}{2}}}{\Gamma(\frac{\nu_1+\nu_2}{2})}dw \\
&= \frac{(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}}{2^{\frac{\nu_1+\nu_2}{2}}\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{(\frac{1}{2}(f\frac{\nu_1}{\nu_2}+1))^{\frac{\nu_1+\nu_2}{2}}} \text{ where } \int_0^\infty \frac{w^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{w}{2}(f\frac{\nu_1}{\nu_2}+1)}(f\frac{\nu_1}{\nu_2}+1))^{\frac{\nu_1+\nu_2}{2}}}{\Gamma(\frac{\nu_1+\nu_2}{2})}dw = 1 \\
&= \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}(f\frac{\nu_1}{\nu_2}+1)^{-\frac{\nu_1+\nu_2}{2}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}(f\frac{\nu_1}{\nu_2}+1)^{-\frac{\nu_1+\nu_2}{2}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}(f\frac{\nu_1}{\nu_2}+1)^{-\frac{\nu_1}{2}-\frac{\nu_2}{2}}
\end{aligned}
$$

1.12.2 Cumulative distribution function

According to (Johnson, Kotz, and Balakrishnan 1995, 2:248) and (Johnson, Kotz, and Balakrishnan 1995, 2:325) we have the following result:

$$
F = \frac{\nu_2}{\nu_1}\frac{X}{1-X} \sim f(\nu_1,\nu_2) \text{ where } X \sim \beta(\frac{\nu_1}{2},\frac{\nu_2}{2})
$$

That is F is distributed as a F-distribution with ν_1 and ν_2 as parameters

We have that $g(x) = \frac{\nu_2}{\nu_1}\frac{x}{1-x}$ is an increasing an continuos function in $x \in (0,1)$ taking into account that $\nu_2 > 0$ and $\nu_1 > 0$. Therefore we have that:

- $g^{-1}(f) = \frac{f\nu_1}{\nu_2+f\nu_1}$
- $g'(f) = \frac{\nu_2}{\nu_1(1-f)^2}$
- $\frac{1}{g'(g^{-1}(f))} = \frac{\frac{1}{1}}{\frac{\frac{\nu_2}{1}}{\nu_1(1-\frac{f\nu_1}{\nu_2+f\nu_1})^2}} = \frac{\frac{1}{1}}{\frac{\frac{\nu_2}{1}}{\nu_1(\frac{\nu_2}{\nu_2+f\nu_1})^2}} = \frac{\nu_1\nu_2}{(\nu_2+f\nu_1)^2}$

So we have the following result:

$$
\begin{aligned}
f_F &= f_X(g^{-1}(f))\frac{1}{g'(g^{-1}(f))} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{f\nu_1}{\nu_2+f\nu_1})^{\frac{\nu_1}{2}-1}(1-\frac{f\nu_1}{\nu_2+f\nu_1})^{\frac{\nu_2}{2}-1}\frac{1}{\nu_1\nu_2}(\frac{\nu_2+f\nu_1}{\nu_2})^2 \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{f\nu_1}{\nu_2+f\nu_1})^{\frac{\nu_1}{2}-1}(\frac{\nu_2}{\nu_2+f\nu_1})^{\frac{\nu_2}{2}-1}\frac{\nu_1\nu_2}{(\nu_2+f\nu_1)^2} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}\nu_1^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}\nu_2^{\frac{\nu_2}{2}}\frac{1}{(\nu_2+f\nu_1)^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}\nu_2^{\frac{\nu_2}{2}}\nu_2^{\frac{\nu_1}{2}}\frac{1}{(\nu_2+f\nu_1)^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}\nu_2^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}\frac{1}{(\nu_2+f\nu_1)^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}\frac{1}{(\frac{\nu_2}{\nu_2}+\frac{f\nu_1}{\nu_2})^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}} \\
&= \frac{1}{\beta(\frac{\nu_1}{2},\frac{\nu_2}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}}f^{\frac{\nu_1}{2}-1}(1+\frac{f\nu_1}{\nu_2})^{-\frac{\nu_1}{2}-\frac{\nu_2}{2}}
\end{aligned}
$$

We can use the above result to find the cumulative function in the following way:

$$
\begin{aligned}
F_F(f) &= Pr[F \leq f] \\
&= Pr[\frac{\nu_2}{\nu_1}\frac{X}{1-X} \leq f]
\end{aligned}
$$

2 Testing Group Frequencies: chisq.test()

2.1 Toy data

2.1.1 Base R

```
data_chisq_test <- rep(c(1:4), times = c(25,25,25,20))
data_chisq_test
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2
[39] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4
[77] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

```
tmp.tab <- table(data_chisq_test)
tmp.tab
```

```
data_chisq_test
 1  2  3  4 
25 25 25 20 
```

2.1.2 Tidyverse

```
tmp_tab <- tibble(x =data_chisq_test)|>
  count(x, name = 'observed') |>
  mutate(expected = (1/4)*sum(observed))
tmp_tab
```

```
# A tibble: 4 x 3
      x observed expected
  <int>    <int>    <dbl>
1     1       25     23.8
2     2       25     23.8
3     3       25     23.8
4     4       20     23.8
```

2.2 chisq.test()

2.2.1 By hand

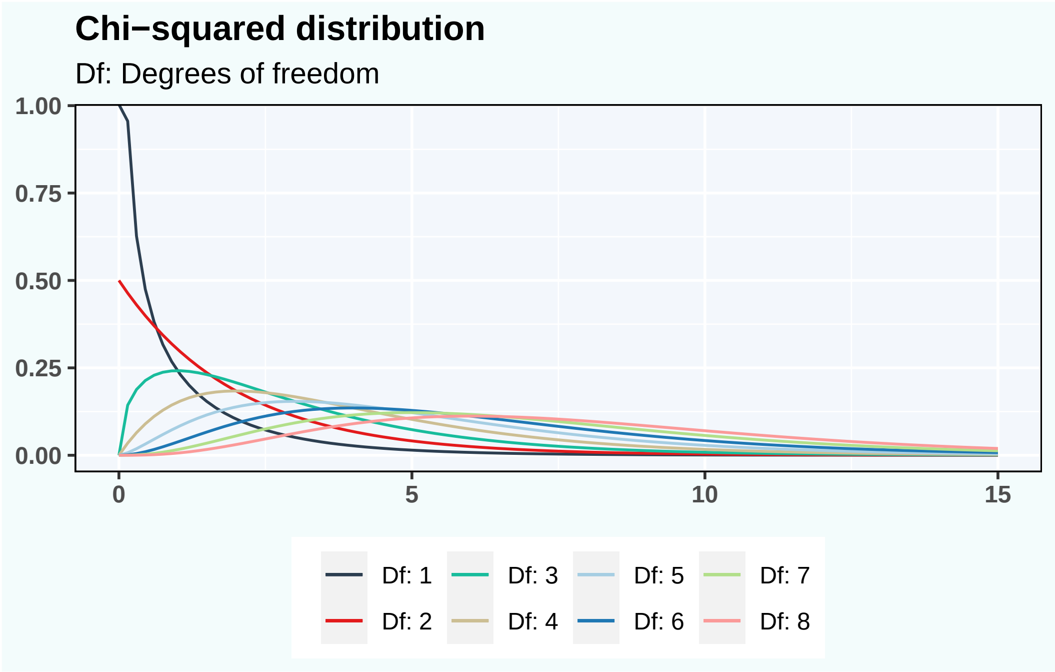
2.2.1.1 Chi-squared distribution: χ^2

$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$ where $Z_i \sim \mathcal{N}(0,1)$ and $Z_1, \dots, Z_k$ are independent random variables

2.2.1.2 Derivation of the chi-squared probability density function

$$
\begin{aligned}
M_{\sum_{i=1}^{k} Z_i^2}(t) &= E[e^{t\sum_{i=1}^{k} Z_i^2}] \\
&= E[e^{t(Z_1^2+\dots+Z_k^2)}] \\
&= E[e^{tZ_1^2} \cdots e^{tZ_k^2}] \\
&= E[e^{tZ_1^2}] \cdots E[e^{tZ_k^2}] \text{ Because } Z_1, \dots Z_k \text{ are independent random variables} \\
&= \frac{1}{(1-2t)^{\frac{1}{2}}} \cdots \frac{1}{(1-2t)^{\frac{1}{2}}} \text{ Because } Z_i^2 \sim \chi^2(1) \\
&= \frac{1}{(1-2t)^{\frac{k}{2}}}
\end{aligned}
$$

So $\sum_{i=1}^{k} Z_i^2$ and $\chi^2(k)$ has the same moment generating function. According to (Rice 2021, chap. 4, page 155) the moment-generating function uniquely determines the probability density function under certain conditions. Using this we can conclude that$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$



2.2.2 Test for given probabilities

$H_0 : p_1 = \frac{1}{4} \wedge p_2 = \frac{1}{4} \wedge p_3 = \frac{1}{4} \wedge p_4 = \frac{1}{4}$

$H_1 : p_1 \neq \frac{1}{4} \vee p_2 \neq \frac{1}{4} \vee p_3 = \frac{1}{4} \vee p_4 \neq \frac{1}{4}$

$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} = \frac{25-95\frac{1}{4}}{95\frac{1}{4}} + \frac{25-95\frac{1}{4}}{95\frac{1}{4}} + \frac{25-95\frac{1}{4}}{95\frac{1}{4}} + \frac{20-95\frac{1}{4}}{95\frac{1}{4}}$

```
chi_square_statistic <- ((tmp_tab$observed - tmp_tab$expected)^2 /
                             tmp_tab$expected) |>
  sum()
chi_square_statistic
```

```
[1] 0.7894737
```

```
degrees_of_freedom <- nrow(tmp_tab) - 1
degrees_of_freedom
```

```
[1] 3
```

```r
p_value <- pchisq(chi_square_statistic,
                  df = degrees_of_freedom,
                  lower.tail = FALSE)
p_value
```

```
[1] 0.8519831
```

p-value $= \mathbf{P}(\chi^2(3) > 0.7894737) = 0.8519831$

Data shows no evidence that the groups in the population are of unequal size, under the assumption of random sampling. In general, a p-value less than 0.10, 0.05 or 0.01 suggests that there is a difference between groups

2.2.2.1 Base R

```r
chisq_test <- chisq.test(x = table(data_chisq_test),
                         p = rep(1/length(tmp.tab),
                                 length(tmp.tab)))
str(chisq_test)
```

```
List of 9
 $ statistic: Named num 0.789
  ..- attr(*, "names")= chr "X-squared"
 $ parameter: Named num 3
  ..- attr(*, "names")= chr "df"
 $ p.value  : num 0.852
 $ method   : chr "Chi-squared test for given probabilities"
 $ data.name: chr "table(data_chisq_test)"
 $ observed : 'table' int [1:4(1d)] 25 25 25 20
  ..- attr(*, "dimnames")=List of 1
  .. ..$ data_chisq_test: chr [1:4] "1" "2" "3" "4"
 $ expected : Named num [1:4] 23.8 23.8 23.8 23.8
  ..- attr(*, "names")= chr [1:4] "1" "2" "3" "4"
 $ residuals: 'table' num [1:4(1d)] 0.256 0.256 0.256 -0.769
  ..- attr(*, "dimnames")=List of 1
  .. ..$ data_chisq_test: chr [1:4] "1" "2" "3" "4"
 $ stdres   : 'table' num [1:4(1d)] 0.296 0.296 0.296 -0.889
  ..- attr(*, "dimnames")=List of 1
  .. ..$ data_chisq_test: chr [1:4] "1" "2" "3" "4"
 - attr(*, "class")= chr "htest"
```

```r
chisq_test$statistic
```

```
X-squared
0.7894737
```

```r
chisq_test$parameter
```

```
df
 3
```

```r
chisq_test$p.value
```

```
[1] 0.8519831
```

```r
chisq_test$method
```

```
[1] "Chi-squared test for given probabilities"
```

```r
chisq_test$observed
```

```
data_chisq_test
 1  2  3  4
25 25 25 20
```

```r
chisq_test$expected
```

```
    1     2     3     4
23.75 23.75 23.75 23.75
```

2.2.2.2 Tidymodels

```r
tibble(x = data_chisq_test) |>
  mutate(x = factor(x, ordered = FALSE)) |>
  chisq_test(response = x,
             p = c(1/4, 1/4, 1/4, 1/4))
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <dbl>   <dbl>
1     0.789        3   0.852
```

2.2.3 Test for independence[5]

```
segmentation <- read_csv(file = '000_data_sets/006_rintro-chapter5.csv')
segmentation |>
  count(subscribe, ownHome) |>
  pivot_wider(id_cols = subscribe,
              names_from = ownHome,
              values_from = n)
```

```
# A tibble: 2 x 3
  subscribe ownNo ownYes
  <chr>     <int>  <int>
1 subNo       137    123
2 subYes       22     18
```

$H_0 : p_{11} = \frac{260}{300}\frac{159}{300} \wedge p_{12} = \frac{260}{300}\frac{141}{300} \wedge p_{21} = \frac{40}{300}\frac{159}{300} \wedge p_{22} = \frac{40}{300}\frac{141}{300}$

$H_1 : p_{11} \neq \frac{260}{300}\frac{159}{300} \vee p_{12} \neq \frac{260}{300}\frac{141}{300} \vee p_{21} \neq \frac{40}{300}\frac{159}{300} \vee p_{22} \neq \frac{40}{300}\frac{141}{300}$

$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} = \frac{(137-300\frac{260}{300}\frac{159}{300})^2}{300\frac{260}{300}\frac{159}{300}} + \frac{(123-300\frac{260}{300}\frac{141}{300})^2}{300\frac{260}{300}\frac{141}{300}} + \frac{(22-300\frac{40}{300}\frac{159}{300})^2}{300\frac{40}{300}\frac{159}{300}} + \frac{(18-300\frac{40}{300}\frac{141}{300})^2}{300\frac{40}{300}\frac{141}{300}}$

3 Testing Observed Proportions: binom.test()

3.1 Toy data

3.1.1 Base R

```
table(segmentation$gender)
```

```

Female   Male 
   157    143 
```

[5] See this resources to understand the calculation of observed and expected values https://online.stat.psu.edu/stat500/lesson/8/8.1

3.1.2 Tidyverse

```r
segmentation |>
  count(gender)
```

```
# A tibble: 2 x 2
  gender      n
  <chr>   <int>
1 Female    157
2 Male      143
```

In our case we define as a success that $gender = Female$ which mean the $Female$ is coded us 1

3.2 binom.test()

3.2.1 By hand

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

$B = \sum_{i=1}^{n} x_i = 157$ where $x_i \in 0, 1$

$\hat{p} = \frac{157}{300} \approx 0.5233$ and $1 - \hat{p} = 1 - \frac{157}{300} \approx 0.4766$

p-value $= \mathbf{P}(X \leq 143) + \mathbf{P}(X \geq 157) \approx 0.2264879 + 0.2264879 \approx 0.4529757$

Using the help of R without using a predefined table we have that:

```r
pbinom(q = 300 - 157,
       size = 300,
       prob = 0.5,
       lower.tail = TRUE) +
pbinom(q = 157 - 1,
       size = 300,
       prob = 0.5,
       lower.tail = FALSE)
```

```
[1] 0.4529757
```

Based on the data at hand we don't have sufficient evidence to reject that $p = 0.5$

In the case of confidence interval for $\alpha = 0.05$ we have the following result:

$$I_x^{-1}\left(p = \frac{0.05}{2}; a = 157, b = 300 - 157 + 1\right) < p < I_x^{-1}\left(p = 1 - \frac{0.05}{2}; a = 157 + 1, b = 300 - 157\right)$$
$$0.4651595 < p < 0.58104184$$

Where $I_x^{-1}(p; a, b)$ is the beta quantile function taking into account that $I_x(a, b)$ is the beta cumulative distribution function. Therefore we have:

$$\begin{aligned} I_x(a, b) &= p \\ x &= I_x^{-1}(p; a, b) \end{aligned}$$

The above interval means that in the long run 95% of confidence intervals constructed in this manner will contain the true parameter

Also this confidence interval is obtained in the following way by taking into account that the interval contains all values of p that are not rejected by the test at confidence level α:

$$\begin{aligned} P_{p_L}(X \geq x) &= \sum_{k=x}^{n} \binom{n}{k} p_L^k (1 - p_L)^{n-k} && = \frac{\alpha}{2} \\ &= I_{p_L}(x, n - x + 1) && = \frac{\alpha}{2} \\ &= p_L && = I_{p_L}^{-1}(\frac{\alpha}{2}; x, n - x + 1) \end{aligned}$$

$$\begin{aligned} P_{p_U}(X \leq x) &= 1 - P_{p_U}(X \geq x + 1) && = \frac{\alpha}{2} \\ &= 1 - I_{p_U}(x + 1, n - (x + 1) + 1) && = \frac{\alpha}{2} \\ &= 1 - I_{p_U}(x + 1, n - x) && = \frac{\alpha}{2} \\ &= I_{p_U}(x + 1, n - x) && = 1 - \frac{\alpha}{2} \\ &= p_U && = I_{p_U}^{-1}(1 - \frac{\alpha}{2}; x + 1, n - x) \end{aligned}$$

3.2.2 Base R

```r
binom_test <- binom.test(x = 157, n = 300, p = 0.5,
           alternative = "two.sided", conf.level = 0.95)
str(binom_test)
```

```
List of 9
 $ statistic  : Named num 157
  ..- attr(*, "names")= chr "number of successes"
 $ parameter  : Named num 300
  ..- attr(*, "names")= chr "number of trials"
 $ p.value    : num 0.453
 $ conf.int   : num [1:2] 0.465 0.581
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num 0.523
  ..- attr(*, "names")= chr "probability of success"
 $ null.value : Named num 0.5
  ..- attr(*, "names")= chr "probability of success"
 $ alternative: chr "two.sided"
 $ method     : chr "Exact binomial test"
 $ data.name  : chr "157 and 300"
 - attr(*, "class")= chr "htest"
```

```r
binom_test$statistic
```

```
number of successes
                157
```

```r
binom_test$estimate
```

```
probability of success
             0.5233333
```

```r
binom_test$p.value
```

```
[1] 0.4529757
```

```
binom_test$conf.int
```

```
[1] 0.4651595 0.5810418
attr(,"conf.level")
[1] 0.95
```

3.2.3 Tidyverse

```
binom_test |>
  tidy()
```

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method     alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>      <chr>
1    0.523       157   0.453       300    0.465     0.581 Exact bin~ two.sided
```

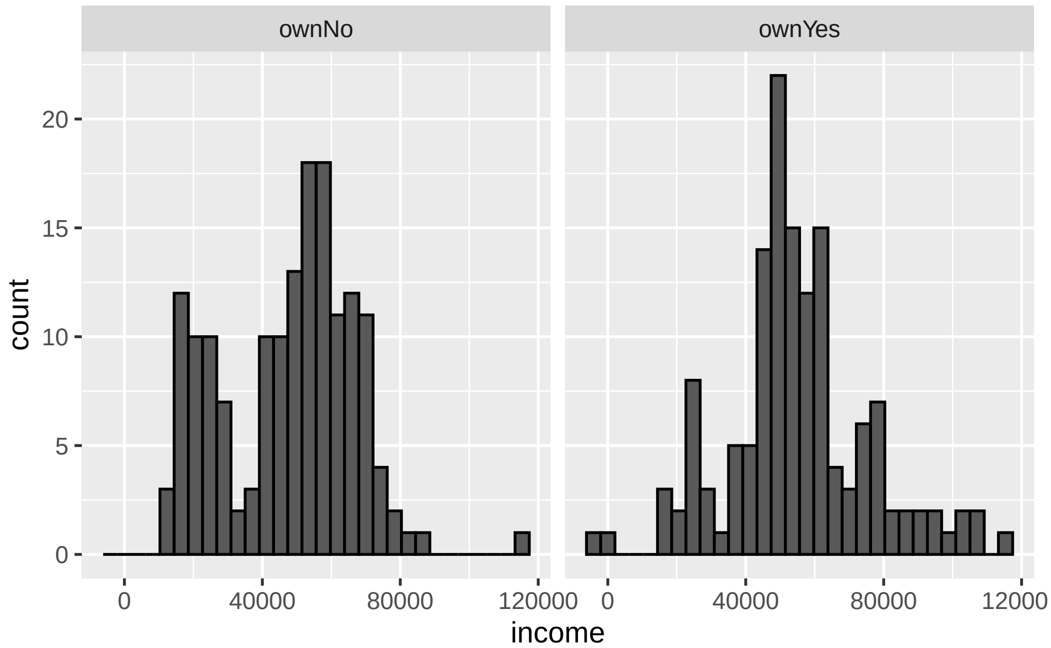
4 Testing Group Means: t.test()

4.1 Toy data

4.1.1 Tidyverse

```
segmentation |>
  ggplot() +
  geom_histogram(aes(x = income),
                 color='black') +
  facet_wrap(facets = vars(ownHome))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
sample_statistics <- segmentation |>
  group_by(ownHome) |>
  summarise(mean_income = mean(income),
                         # sample variance
                         ## use n - 1 in the denominator
            var_income = var(income),
            n = n())
sample_statistics
```

```
# A tibble: 2 x 4
  ownHome mean_income var_income     n
  <chr>         <dbl>      <dbl> <int>
1 ownNo        47391. 358692875.   159
2 ownYes       54935. 430890091.   141
```

4.2 t.test()

4.2.1 By hand

$H_0 : \mu_{ownNo} - \mu_{ownYes} = 0$

$H_1 : \mu_{ownNo} - \mu_{ownYes} \neq 0$

$t = \frac{\overline{ownNo} - \overline{ownYes}}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} + \frac{s^2_{ownYes}}{n_{ownYes}}}} = \frac{47391.01 - 54934.68}{\sqrt{\frac{358692875}{159} + \frac{430890091}{141}}} \approx -3.273094$

- Degrees of freedom

$\nu \approx \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{(\frac{s_1^2}{N_1})^2}{N_1 - 1} + \frac{(\frac{s_2^2}{N_2})^2}{N_2 - 1}} = \frac{(\frac{358692875}{159} + \frac{430890091}{141})^2}{\frac{(\frac{358692875}{159})^2}{159-1} + \frac{(\frac{430890091}{141})^2}{141-1}} = 285.2521$

Using R we have the following result:

```
((sample_statistics$var_income[1] / sample_statistics$n[1]) +
   (sample_statistics$var_income[2] / sample_statistics$n[2]))^2 /
  ((sample_statistics$var_income[1] / sample_statistics$n[1])^2 /
     (sample_statistics$n[1] - 1) + (sample_statistics$var_income[2] /
     sample_statistics$n[2])^2 /
     (sample_statistics$n[2] - 1))
```

```
[1] 285.2521
```

p-value $= \mathbf{P}(T \leq t) + \mathbf{P}(T \geq t) \approx 0.0005973553 + 0.0005973553 \approx 0.001194711$

- Confidence interval

$$P\left(t_L < \frac{\overline{x}_{ownNo} - \overline{x}_{ownYes} - (\mu_{ownNo} - \mu_{ownYes})}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} + \frac{s^2_{ownYes}}{n_{ownYes}}}} < t_U\right) = 0.95$$

We need to specify t_L and t_U:

```
t_L <- qt(p = 0.025, df = 285.25, lower.tail = TRUE)
t_L
```

```
[1] -1.968315
```

```
t_U <- qt(p = 0.975, df = 285.25, lower.tail = TRUE)
t_U
```

```
[1] 1.968315
```

Therefore we have that:

$$P\left(-t_{0.025,\nu} < \frac{\overline{x}_{ownNo} - \overline{x}_{ownYes} - (\mu_{ownNo} - \mu_{ownYes})}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} + \frac{s^2_{ownYes}}{n_{ownYes}}}} < t_{0.025,\nu}\right) = 0.95$$

Where $\nu \approx \frac{(\frac{s^2_{ownNo}}{n_{ownNo}} + \frac{s^2_2}{n_{ownYes}})^2}{\frac{(\frac{s^2_{ownNo}}{n_{ownNo}})^2}{n_{ownNo}-1} + \frac{(\frac{s^2_2}{n_{ownYes}})^2}{n_{ownYes}-1}} = 285.2521$

$$P(-7543.674 - 1.968315 \times 2304.753 < \mu_{ownNo} - \mu_{ownYes} < -7543.674 - 1.968315 \times 2304.753) = 0.95$$

$$P(-12080.16 < \mu_{ownNo} - \mu_{ownYes} < -3007.193) = 0.95$$

- In the long run 95% of confidence intervals constructed in this manner will contain the true parameter

4.2.2 Base R

```
t_test <- t.test(income ~ ownHome, data = segmentation,
                 alternative ='two.sided', mu = 0,
                 # See https://www.statology.org/paired-vs-unpaired-t-test/
                 paired = FALSE,
                 # Welch's t-test
                 var.equal = FALSE,
                 conf.level = 0.95)
str(t_test)
```

```
List of 10
 $ statistic  : Named num -3.27
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 285
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.00119
 $ conf.int   : num [1:2] -12080 -3007
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num [1:2] 47391 54935
  ..- attr(*, "names")= chr [1:2] "mean in group ownNo" "mean in group ownYes"
 $ null.value : Named num 0
```

```
  ..- attr(*, "names")= chr "difference in means between group ownNo and group ownYes"
 $ stderr     : num 2305
 $ alternative: chr "two.sided"
 $ method     : chr "Welch Two Sample t-test"
 $ data.name  : chr "income by ownHome"
 - attr(*, "class")= chr "htest"
```

4.2.3 Tidyverse

```r
segmentation |>
  t_test(formula = income ~ ownHome,
         alternative = "two-sided",
         order = c("ownNo", "ownYes"),
         mu = 0,
         conf_level = 0.95)
```

```
# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
1     -3.27  285. 0.00119 two.sided     -7544.  -12080.   -3007.
```

5 Testing Multiple Group Means: aov() and anova()

5.1 Toy data

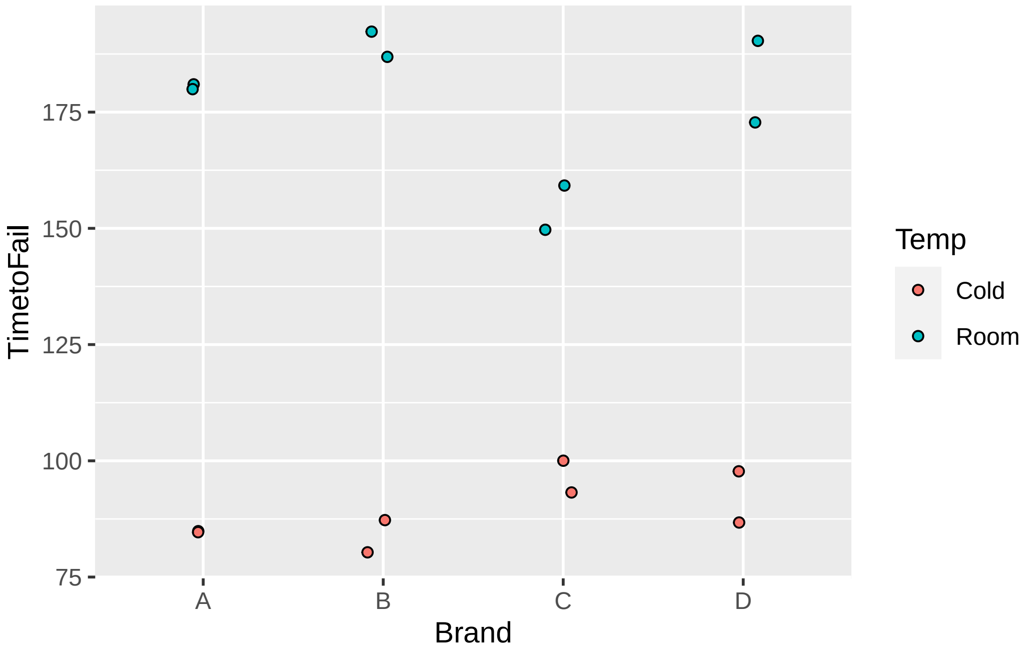
5.1.1 Tidyverse

```r
toy_data <- tribble(
  ~TimetoFail,  ~Temp, ~Brand, ~Group,
         181L, "Room",    "A",   "RA",
         187L, "Room",    "B",   "RB",
         150L, "Room",    "C",   "RC",
         173L, "Room",    "D",   "RD",
          85L, "Cold",    "A",   "CA",
          80L, "Cold",    "B",   "CB",
          93L, "Cold",    "C",   "CC",
          87L, "Cold",    "D",   "CD",
         180L, "Room",    "A",   "RA",
```

```r
         192L, "Room",     "B",    "RB",
         159L, "Room",     "C",    "RC",
         190L, "Room",     "D",    "RD",
          85L, "Cold",     "A",    "CA",
          87L, "Cold",     "B",    "CB",
         100L, "Cold",     "C",    "CC",
          98L, "Cold",     "D",    "CD"
  )
toy_data |>
  arrange(Temp, Brand)
```

```
# A tibble: 16 x 4
   TimetoFail Temp  Brand Group
        <int> <chr> <chr> <chr>
 1         85 Cold  A     CA
 2         85 Cold  A     CA
 3         80 Cold  B     CB
 4         87 Cold  B     CB
 5         93 Cold  C     CC
 6        100 Cold  C     CC
 7         87 Cold  D     CD
 8         98 Cold  D     CD
 9        181 Room  A     RA
10        180 Room  A     RA
11        187 Room  B     RB
12        192 Room  B     RB
13        150 Room  C     RC
14        159 Room  C     RC
15        173 Room  D     RD
16        190 Room  D     RD
```

```r
toy_data |>
  ggplot(aes(x = Brand, y = TimetoFail, fill = Temp)) +
  geom_point(position = position_jitter(width = 0.1),
             shape=21, color='black')
```



5.2 One way Anova: aov() & anova()

In the case of one way anova it is assumed that there are many y_{ij} with $i = 1, \ldots, n_j$ and $j = 1, \ldots, a$ where $n = \sum_{j=1}^{a} n_j$

For example in the case of `toy_data` with the variables `TimetoFail` and `Temp` we have that $j = 1, 2$, $n_1 = 8$, $n_2 = 8$:

```
toy_data |>
  group_by(Temp) |>
  summarize(n = n())
```

```
# A tibble: 2 x 2
  Temp      n
  <chr> <int>
1 Cold      8
2 Room      8
```

Also it is assumed that each y_{ij} is a realization of a random variable:

$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

Specifically it is assumed the following:

$$
\begin{aligned}
Y_{ij} &= \mu_j + \epsilon_{ij} \\
&= \mu + \beta_j + \epsilon_{ij}
\end{aligned}
$$

Where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\mu_i = \mu + \beta_j$. In the case of the variables `TimetoFail` and `Temp` we need to estimate the following parameters: $\mu_1, \mu_2, \mu, \sigma^2$. Therefore we have $a+1$ parameters to estimate because σ^2 can be estimated using to estimations of $\mu_1, \ldots, \mu_a, \mu$.

In `R` the following constrain are imposed using `contr.treatment`:

$$
\beta_1 = 0 \text{ and } \mu = \mu_1
$$

The problem to solve is the following:

$$
\hat{\mu}, \hat{\beta}_j = \arg\min_{\mu, \beta_j} \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \mu - \beta_j)^2
$$

In the literature the following notation is used:

- $y_{.j} = \sum_{i=1}^{n_j} y_{ij}$ sum of group j
- $y_{..} = \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij}$ sum of all observations
- $\overline{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ mean of group j
- $\overline{y}_{..} = \frac{1}{n} \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij}$ total mean

Solving the problem we have that:

$$
\begin{aligned}
\sum_{i=1}^{n_j} -2(y_{ij} - \hat{\mu} - \hat{\beta}_j) = 0 \\
\sum_{i=1}^{n_j} y_{ij} &= n_j(\hat{\beta}_j + \hat{\mu}) \\
\overline{y}_{.j} - \hat{\mu} &= \hat{\beta}_j
\end{aligned}
$$

Also because $\beta_1 = 0$ we have that $\overline{y}_{.1} - \hat{\mu} = 0$ so $\hat{\mu} = \overline{y}_{.1}$. Therefore we have that:

- $\hat{\mu} = \overline{y}_{.1}$
- $\hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{.1}$ for $j = 2, \ldots, a$
- $\hat{\mu}_j = \overline{y}_{.j}$ for $j = 1, \ldots, a$

Finally we have that $\hat{\sigma}^2$ is given by:

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-a}\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\hat{\mu}_j)^2 \\
&= \frac{1}{n-a}\sum_{j=1}^{a}(n_j-1)\sum_{i=1}^{n_j}\frac{(y_{ij}-\hat{\mu}_j)^2}{n_j-1} \\
&= \frac{1}{n-a}\sum_{j=1}^{a}(n_j-1)s_j^2
\end{aligned}
$$

5.2.1 Decomposition of variance

- SST: Sum of Squares Total
- SSB: Sum of Squares Between
- SSW: Sum of Squares Within

$$
\begin{aligned}
SST &= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{..})^2 \\
&= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j}+\overline{y}_{.j}-\overline{y}_{..})^2 \\
&= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2+2(y_{ij}-\overline{y}_{.j})(\overline{y}_{.j}-\overline{y}_{..})+(\overline{y}_{.j}-\overline{y}_{..})^2 \\
&= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(\overline{y}_{.j}-\overline{y}_{..})^2+\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2+2\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})(\overline{y}_{.j}-\overline{y}_{..})
\end{aligned}
$$

We have the following result:

$$
\begin{aligned}
\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})(\overline{y}_{.j}-\overline{y}_{..}) &= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}\overline{y}_{.j}-y_{ij}\overline{y}_{..}-\overline{y}_{.j}^2+\overline{y}_{.j}\overline{y}_{..}) \\
&= \sum_{j=1}^{a}n_j\overline{y}_{.j}^2-n\overline{y}_{..}^2-\sum_{j=1}^{a}n_j\overline{y}_{.j}^2+n_j\overline{y}_{.j}^2 \\
&= 0
\end{aligned}
$$

Therefore we have the following result:

$$\begin{aligned} SST &= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(\overline{y}_{.j}-\overline{y}_{..})^2 + \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2 + 2\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})(\overline{y}_{.j}-\overline{y}_{..}) \\ &= \sum_{j=1}^{a}\sum_{i=1}^{n_j}(\overline{y}_{.j}-\overline{y}_{..})^2 + \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2 \\ &= SSB + SSW \end{aligned}$$

Now we can define the following hypothesis:

$H_0 : \mu_{Room} = \mu_{Cold}$

H_1 : At least one group mean is different from the rest

$n_j = [8, 8]$

$\overline{y}_{.j} = [89.375, 176.500]$

$\overline{y}_{..} = 132.9375$

$F_{1,14} = \frac{\frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(\overline{y}_{.j}-\overline{y}_{..})^2}{a-1}}{\frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2}{n-a}} = \frac{\frac{\sum_{j=1}^{2} n_j(\overline{y}_{.j}-\overline{y}_{..})^2}{2-1}}{\frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j}(y_{ij}-\overline{y}_{.j})^2}{16-2}} = \frac{\frac{30363.06}{1}}{\frac{1923.875}{14}} = 220.9514$

$Pr[F_{1,14} > 220.9514] = 0.0000000005738688$

$F^{-1}_{F_{1,14}}(p = 0.95) = 4.60011 < 220.9514$

5.3 Two way Anova: aov() & anova()

Becareful: Read - https://cran.r-project.org/doc/FAQ/R-FAQ.html#Why-does-the-output-from-anova_0028_0029-depend-on-the-order-of-factors-in-the-model_003f

- https://www.r-bloggers.com/2011/03/anova-%E2%80%93-type-iiiiii-ss-explained/
-

We want to explain the variance of the data. First we define the following matrix $\mathbf{M}_{\mathbf{1}_n} \equiv \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n^T\mathbf{1}_n)^{-1}\mathbf{1}_n^T$. In general we can define $\mathbf{M}_X \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. So we have the following result:

$$
\mathbf{M}_{\mathbf{1}_n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n} - \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}_{n \times n}
$$

$$
= \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n} \end{bmatrix}_{n \times n}
$$

```r
dim_toy_data <- dim(toy_data)
n_row <- dim_toy_data[1]
one_n_row <- model.matrix(object = TimetoFail ~ 1, data=toy_data)
M_1_n <- (diag(x = n_row) - one_n_row %*% solve(t(one_n_row) %*% one_n_row)
         %*% t(one_n_row))
M_1_n
```

```
         1       2       3       4       5       6       7       8       9
1   0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
2  -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
3  -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
4  -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
5  -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625
6  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625
7  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625
8  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625
9  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375
10 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
11 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
12 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
13 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
14 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
15 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
16 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
        10      11      12      13      14      15      16
1  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
2  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
3  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
4  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
5  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
6  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
```

```
7  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
8  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
9  -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
10  0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
11 -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
12 -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625 -0.0625
13 -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625 -0.0625
14 -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625 -0.0625
15 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375 -0.0625
16 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.9375
```

References

Johnson, Norman Lloyd, Adrienne W. Kemp, and Samuel Kotz. 2005. *Univariate Discrete Distributions*. 3rd ed. Hoboken, N.J.: Wiley-Interscience.

Johnson, Norman Lloyd, Samuel Kotz, and Narayanaswamy Balakrishnan. 1995. *Continuous Univariate Distributions*. 2nd ed. Vol. 2. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Rice, John A. 2021. *Mathematical Statistics and Data Analysis*. 3rd ed., 9th Indian reprint. New Delhi: Cengage Learning.