STAT 500

Several Useful Distributions

Assumptions for Model-based Inference

- It is commonly assumed that:
  - $Y_{11}, \ldots, Y_{1n_1}$ are iid $N(\mu_1, \sigma^2)$ random variables
  - $Y_{21}, \ldots, Y_{2n_2}$ are iid $N(\mu_2, \sigma^2)$ random variables
  - The samples are independent
  - Note the homogeneous variance assumption

- This is equivalent to the linear model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

  with the ϵ_{ij} as iid $N(0, \sigma^2)$ random variables

  $i = 1, 2$ treatment groups

  $j = 1, 2, ..., n_i$ units in the i-th group

The Normal Distribution

Definition: A random variable $\boldsymbol{Y}$ with density function

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

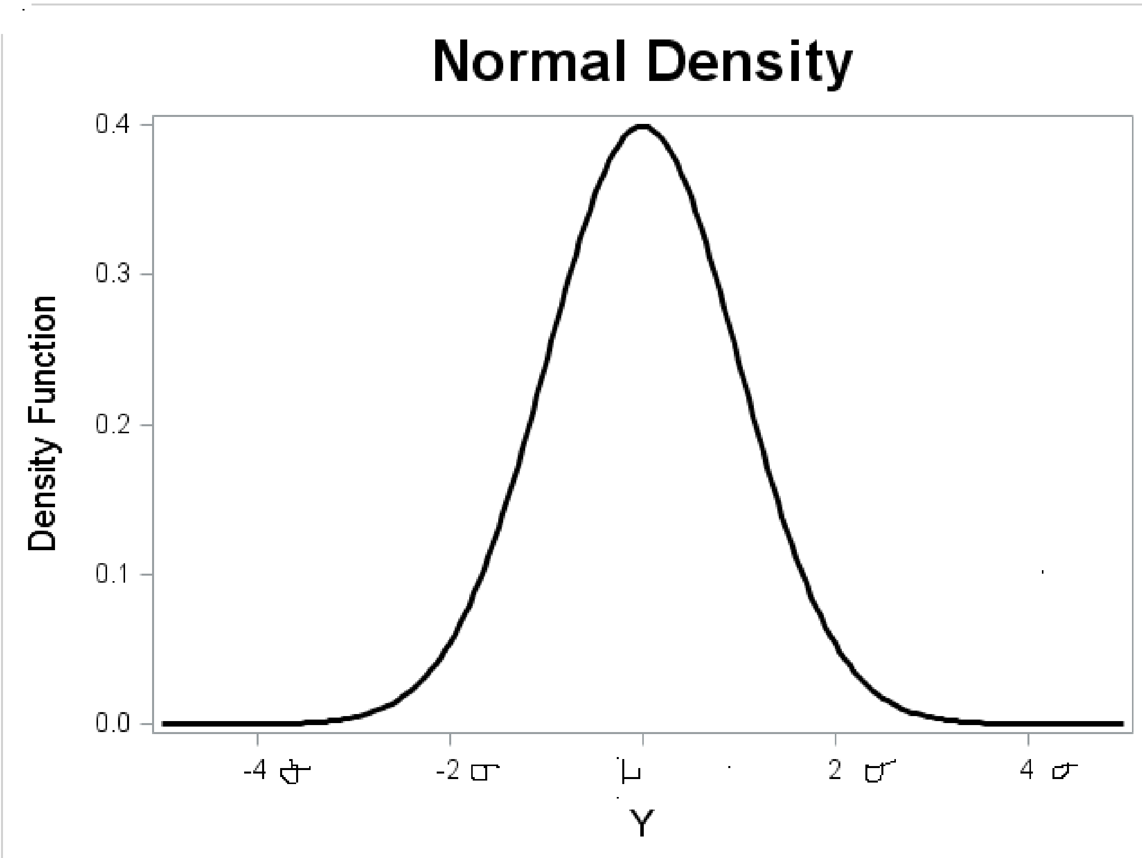
is said to have a normal (*Gaussian*) distribution with

$$\text{Mean} \equiv \mathrm{E}(Y) = \mu \quad \text{and} \quad \text{Variance} \equiv \mathrm{Var}(Y) = \sigma^2$$

The *standard deviation* is $\sigma = \sqrt{\mathrm{Var}(Y)}$

We will use the notation $Y \sim N(\mu, \sigma^2)$

Normal Distribution



Standard Normal Distribution

Suppose Z has a normal distribution with $E(Z) = 0$ and $Var(Z) = 1$, i.e.,

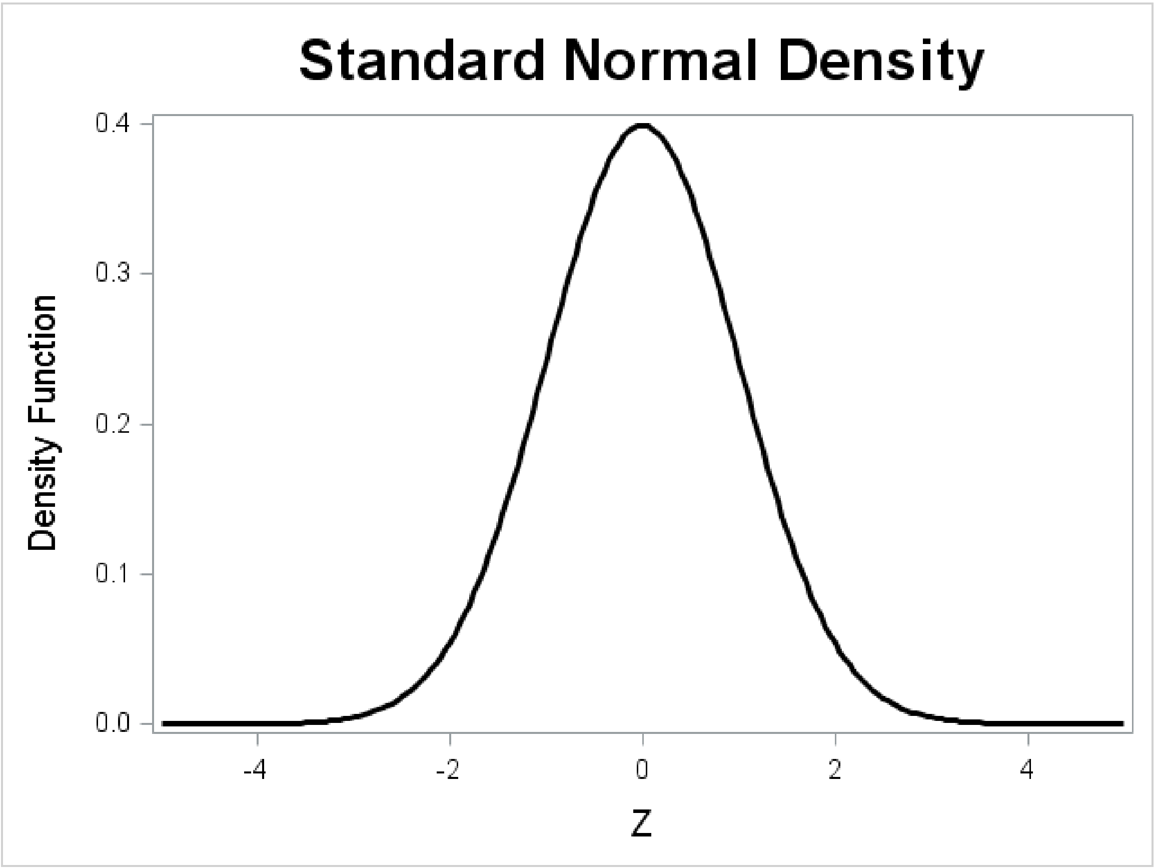
$$Z \sim N(0, 1),$$

then Z is a random variable with a *standard normal* distribution.

If $Z \sim N(0, 1)$ then $Y = (\sigma Z + \mu) \sim N(\mu,\ \sigma^2)$

If $Y \sim N(\mu,\ \sigma^2)$ then $Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$

Standard Normal Distribution



Linear Combinations of Random Variables

If Y_1 is a random variable with expectation μ_1 and variance σ_1^2 and Y_2 is a random variable with expectation μ_2 and variance σ_2^2, then

- $E(Y_1 + Y_2) = \mu_1 + \mu_2$
- $E(aY_1 + bY_2 + c) = a\mu_1 + b\mu_2 + c$
- $Var(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2$ if Y_1 and Y_2 are independent
- $Var(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2$ if Y_1 and Y_2 are independent
- $Var(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 + 2Cov(Y_1, Y_2)$
- $Var(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2abCov(Y_1, Y_2)$

Linear Combinations of Random Variables

Variance: $Var(Y_1) = \sigma_1^2 = E[(Y_1 - \mu_1)^2]$

Covariance: $Cov(Y_1,\ Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = \rho_{12}\sigma_1\sigma_2$
where ρ_{12} is the correlation between Y_1 and Y_2

The correlation coefficient ρ_{12} measures the strength of the linear relationship between Y_1 and Y_2.

Note that $\rho_{12} = \frac{Cov(Y_1,\ Y_2)}{\sigma_1\sigma_2}$ is unit free and

- Always between -1 and 1
- Zero when there is no linear association
- Zero if Y_1 and Y_2 are independent of each other

Linear Combinations of Independent Normal Random Variables

If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$
and Y_1 is <u>independent</u> of Y_2 then

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

and

$$aY_1 + bY_2 + c \sim N(a\mu_1 + b\mu_2 + c,\ a^2\sigma_1^2 + b^2\sigma_2^2)$$

A special case of the second result with $a = 1$ and $b = -1$ yields

$$Y_1 - Y_2 \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

Distribution of a Sample Mean

- Suppose $Y_{k1}, \ldots, Y_{kn_k}$ denotes a simple random sample of n_k observations from a population with population mean μ_k and variance σ_k^2

- $Y_{k1}, \ldots, Y_{kn_k}$ are iid random variables, each with mean μ_k and variance σ_k^2

- The sample mean, $\bar{Y}_k = \sum_{j=1}^{n_k} Y_{k,j}/n_k$, is a random variable with expectation

$$
\begin{aligned}
E(\bar{Y}_k) &= E\left(\frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j}\right) \\
&= \frac{1}{n_k}\sum_{j=1}^{n_k} E\left(Y_{k,j}\right) = \frac{1}{n_k}\sum_{j=1}^{n_k}\mu_k = \mu_k
\end{aligned}
$$

Distribution of a Sample Mean

- The variance of the k-th sample mean is

$$
\begin{aligned}
Var(\bar{Y}_k) &= Var\left(\frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j}\right) \\
&= \frac{1}{n_k^2} Var\left(\sum_{j=1}^{n_k} Y_{k,j}\right) \\
&= \frac{1}{n_k^2}\sum_{j=1}^{n_k} Var\left(Y_{k,j}\right) \\
&= \frac{1}{n_k^2}\sum_{j=1}^{n_k} \sigma_k^2 = \frac{\sigma_k^2}{n_k}
\end{aligned}
$$

Distribution of a Sample Mean

- Assuming independent observations from a population with mean μ_k, the sample mean $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j}$ is the best linear unbiased estimator for μ_k

- If $Y_{k,1}, Y_{k,2}, \cdots, Y_{k,n_k}$ are iid $N(\mu_k,\ \sigma_k^2)$ random variables, i.e., a simple random sample from a normal population, then

$$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j} \ \sim\ N\left(\mu_k, \frac{\sigma_k^2}{n_k}\right)$$

- $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j}$ is a random variable (an *estimator*).
  We will use $\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{k,j}$ to denote an *estimate* of the population mean, an *observed value* of $\bar{Y}_k$.

Distribution of the Difference between Two Sample Means

Independent simple random samples from two normal populations

$$Y_{11}, \ldots, Y_{1n_1} \text{ are } iid\ N(\mu_1, \sigma_1^2) \text{ random variables}$$
$$Y_{21}, \ldots, Y_{2n_2} \text{ are } iid\ N(\mu_2, \sigma_2^2) \text{ random variables}$$

We can derive that: $\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

- To draw inference on $\mu_1 - \mu_2$, we would need to know σ_1^2 and σ_2^2.
- σ_1^2 and σ_2^2 are population parameters and are generally unknown.

Estimation for Variances

$$S_1^2 = \frac{1}{n_1-1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$$

is an unbiased estimator for Var$(Y_{1j}) = \sigma_1^2$ with $n_1 - 1$ degrees of freedom

$$S_2^2 = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$$

is an unbiased estimator for Var$(Y_{2j}) = \sigma_2^2$ with $n_2 - 1$ degrees of freedom

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

is a pooled estimator for σ^2, with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom, when $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Estimation for Variances

When $\boldsymbol{\sigma_1^2 \neq \sigma_2^2}$ estimate Var$(\bar{\boldsymbol{Y}}_{\mathbf{1}} - \bar{\boldsymbol{Y}}_{\mathbf{2}}) = \dfrac{\boldsymbol{\sigma_1^2}}{\boldsymbol{n_1}} + \dfrac{\boldsymbol{\sigma_2^2}}{\boldsymbol{n_2}}$ as

$$\frac{\boldsymbol{S_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{S_2^2}}{\boldsymbol{n_2}}$$

When $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$ estimate Var$(\bar{\boldsymbol{Y}}_{\mathbf{1}} - \bar{\boldsymbol{Y}}_{\mathbf{2}}) = \boldsymbol{\sigma^2}\left(\dfrac{\mathbf{1}}{\boldsymbol{n_1}} + \dfrac{\mathbf{1}}{\boldsymbol{n_2}}\right)$ as

$$\boldsymbol{S_P^2}\left(\frac{\mathbf{1}}{\boldsymbol{n_1}} + \frac{\mathbf{1}}{\boldsymbol{n_2}}\right)$$

Model-based Inference

- Assume each sample is a simple random sample from a population with a normal distribution, the samples are independent, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

- It follows that $$\frac{(n_1 + n_2 - 2)\, S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

  where $\chi^2_{n_1+n_2-2}$ is a central chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom.

Central Chi-Square Distribution

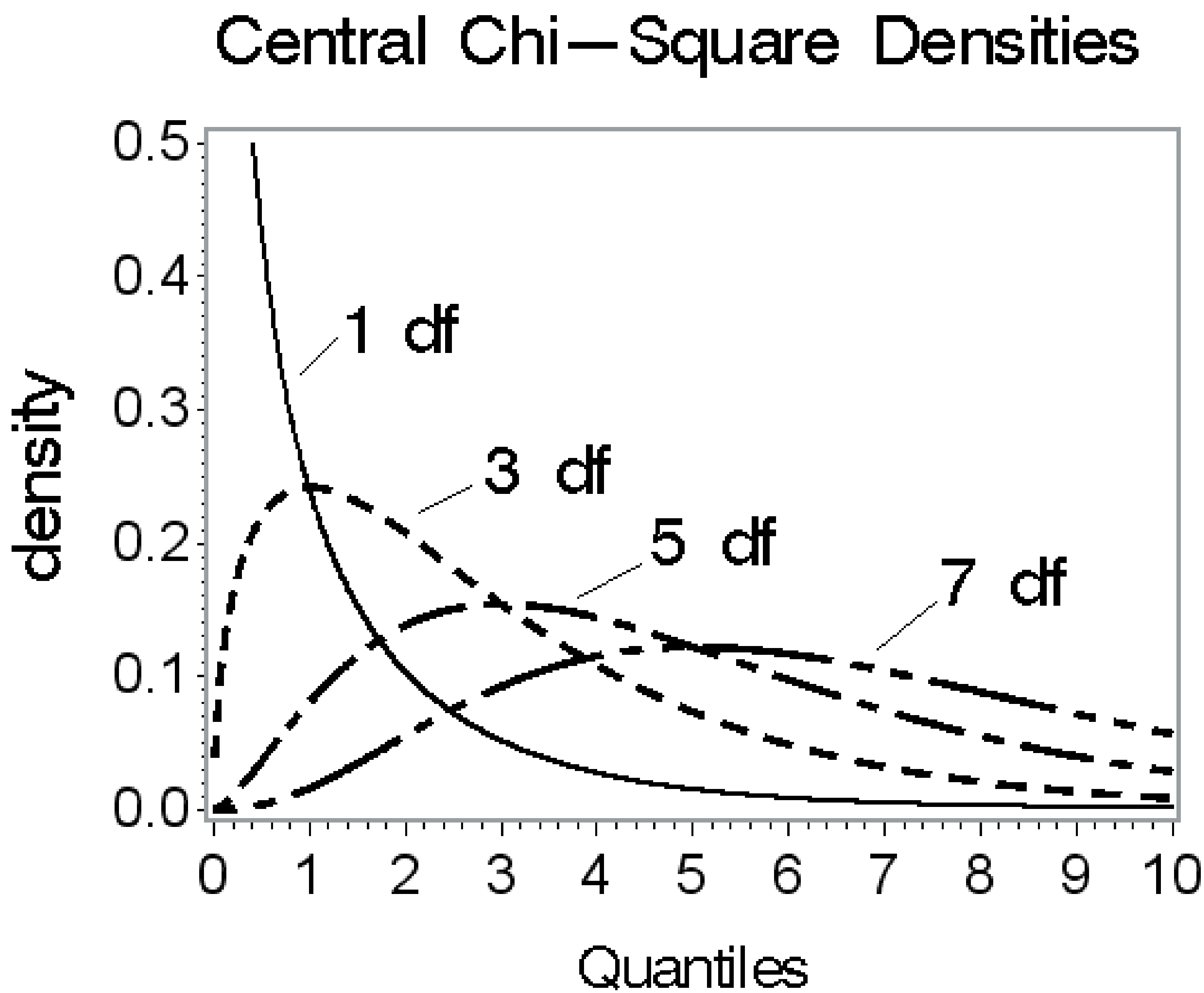
Defn: Let $Z_i, i = 1, 2, \cdots, n,$ be independent standard normal random variables. The distribution of

$$W = \sum_{i=1}^{n} Z_i^2$$

is called the *central chi-square distribution* with n degrees of freedom.

To indicate that a random variable has a central chi-square distribution with ν degrees of freedom, we will use the notation

$$W \sim \chi^2_{(\nu)}$$


Central Chi−Square Densities
density
Quantiles
1 df
3 df
5 df
7 df
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
9
10

Central Chi-Square Distribution

- It can be shown that

$$\frac{(n_i - 1)S_i^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

  is the sum of the squares of $n_i - 1$ independent standard normal random variables

- Consequently, $\dfrac{(n_i - 1)S_i^2}{\sigma^2}$ has a central chi-square distribution with $n_i - 1$ df

Degrees of Freedom (d.f.)

- Quantify the amount of information available to estimate a population variance.
- Consider the sample variance

$$S_i^2 = \Sigma_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n_i - 1) \text{ with } n_i - 1 \text{ d.f.}$$

- Start with n_i observations
  Estimate the population mean with $\bar{Y}_i$
  This imposes one linear restriction $\Sigma_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) = 0$
- Consequently, the vector of residuals

$$\mathrm{e}_i = (Y_{i1} - \bar{Y}_i,\ Y_{i2} - \bar{Y}_i, \ldots, Y_{in_i} - \bar{Y}_i)^T$$

  is an n_i dimensional vector that is restricted to lie in an $n_i - 1$ dimensional linear sub-space.

Sum of Independent Chi-Square Random Variables

- The sum of two independent central chi-square random variables with v_1 and v_2 df, respectively, has a central chi-square distribution with $v_1 + v_2$ df

- Consequently,

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

  has a central chi-square distribution with $(n_1 - 1) + (n_2 - 1)$ df

Student t-distribution

For the model assumptions

- Two independent random samples
- Homogeneous population variances
- Normality

It follows that $Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$ and

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where $t_{n_1+n_2-2}$ denotes a central Student t distribution with $n_1 + n_2 - 2$ d.f.

Student t-distribution

Defn: If $Z \sim N(0,1)$ and $W \sim \chi^2_{(r)}$ and Z and W are independent random variables, then the random variable

$$T = \frac{Z}{\sqrt{W/r}}$$

has a central Student t-distribution with r d.f.

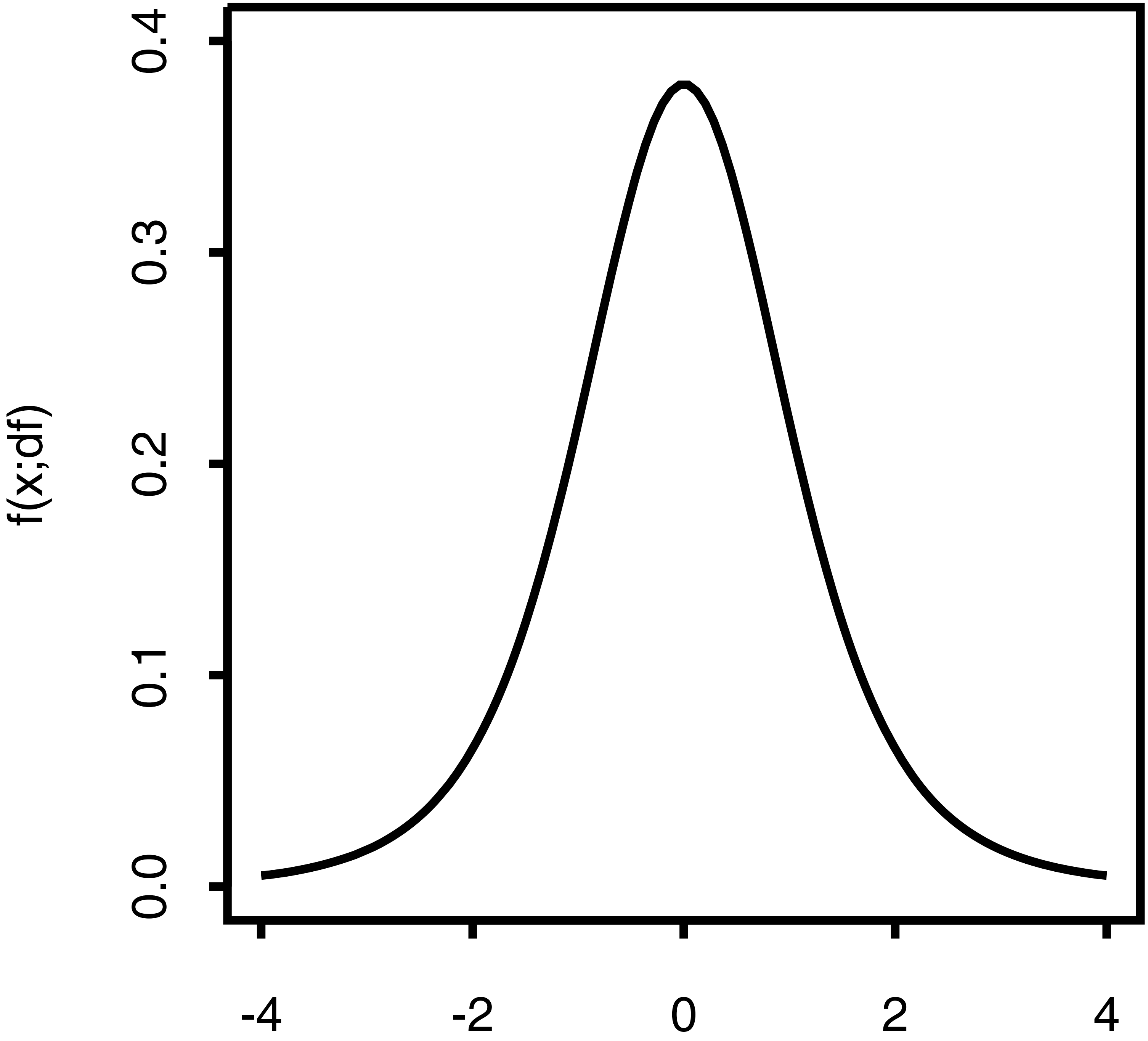
To indicate that a random variable has a central t-distribution with r degrees of freedom, we will use the notation

$$T \sim t_r$$

Properties of the Central t-distribution

- Centered at zero (mean and median are zero)
- Symmetric distribution
- Thicker tails than the standard normal distribution
- Approaches the standard normal distribution as the degrees of freedom become larger
- t_∞ is the standard normal distribution
- 97.5 percentile is around 2 except for small d.f. (e.g. 2.571 for 5 d.f., 2.093 for 19 d.f., 2.000 for 60 d.f., 1.96 for ∞ d.f.)

Central t Density
(5 degrees of freedom)



Central t Densities