t-tests

Recap: t distribution

If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

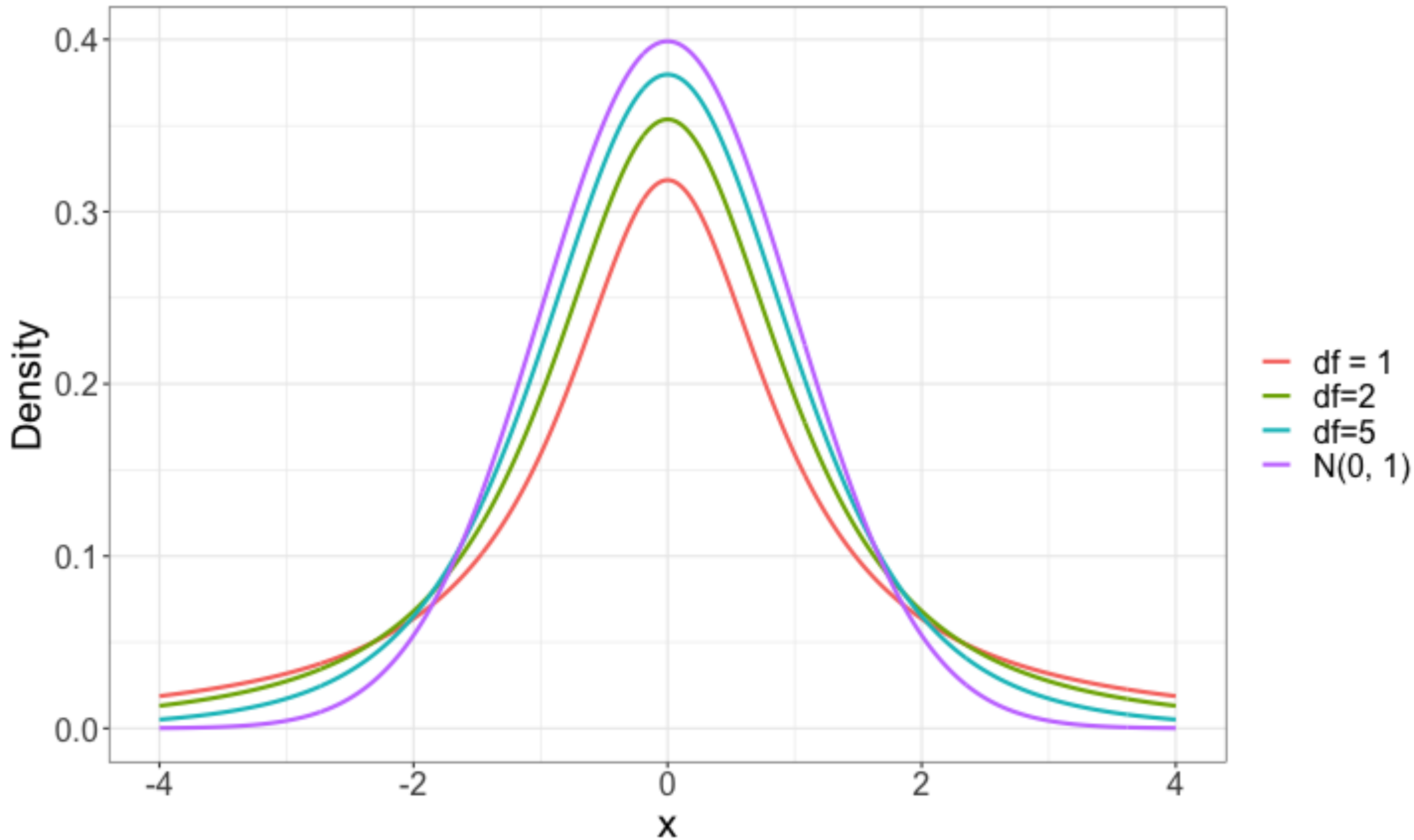
$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \sim t_{n-1}$$

Definition: Let $Z \sim N(0,1)$ and $V \sim \chi^2_d$ be independent. Then

$$T = \frac{Z}{\sqrt{V/d}} \sim t_d$$

As $n \to \infty$, $t_n \approx N(0,1)$

t-distribution



Intuition: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S}$ is more variable than $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$

As $n \to \infty$, $S \xrightarrow{P} \sigma$, so $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \approx \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$

Derivation: WTS if $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$

① $$\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \underbrace{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}_{N(0,1)} \cdot \sqrt{\frac{(n-1)\sigma^2}{(n-1)S^2}}$$

$\Rightarrow$ WTS $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ and WTS $\frac{(n-1)S^2}{\sigma^2} \perp\!\!\!\perp \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$

② $$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

we know $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$ (use $\sum_i (X_i - \bar{X}) = 0$)

Now, $$\underbrace{\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\chi^2_n} = \underbrace{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{\chi^2_{n-1}?} + \underbrace{n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2}_{\chi^2_1?}$$

Cochran's theorem

Let $Z_1, \dots, Z_n \overset{iid}{\sim} N(0,1)$, and let $Z = [Z_1, \dots, Z_n]^T$. Let $A_1, \dots, A_k \in \mathbb{R}^{n \times n}$ be symmetric matrices such that $Z^T Z = \sum_{i=1}^{k} Z^T A_i Z$, and let $r_i = rank(A_i)$. Then the following are equivalent:

- $r_1 + \cdots + r_k = n$
- The $Z^T A_i Z$ are independent
- Each $Z^T A_i Z \sim \chi^2_{r_i}$

Application to t-tests

Let $z_i = \frac{x_i - \mu}{\sigma}$, $Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$ $\qquad \sum_i z_i^2 = Z^T Z$

$$\Rightarrow Z^T Z = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma}\right)^2 + n\left(\frac{\bar{x} - \mu}{\sigma}\right)^2$$

want to find A_1, A_2 such that $Z^T A_1 Z = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma}\right)^2$

and $Z^T A_2 Z = n\left(\frac{\bar{x} - \mu}{\sigma}\right)^2$

$$\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \sum_{i=1}^n (z_i - \bar{z})^2 = \sum_{i=1}^n \left(z_i - \frac{1}{n}\sum_j z_j\right)^2$$

$$n\left(\frac{\bar{x} - \mu}{\sigma}\right)^2 = n(\bar{z})^2 = \frac{1}{n}\left(\sum_j z_j\right)^2$$

$$= \frac{1}{n}[z_1 \cdots z_n] \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

$$= Z^T\left(\frac{1}{n} J_n\right) Z \qquad J_n = n \times n \text{ matrix of all 1s}$$

$A_2 = \frac{1}{n} J_n \qquad \text{rank}(A_2) = 1$

$$Z^T Z = \sum_i \left(Z_i - \tfrac{1}{n}\sum_j Z_j\right)^2 + Z^T\left(\tfrac{1}{n}J_n\right)Z$$

$$\Rightarrow \sum_i \left(Z_i - \tfrac{1}{n}\sum_j Z_j\right)^2 = Z^T Z - Z^T\left(\tfrac{1}{n}J_n\right)Z$$

$$= Z^T\left(I_n - \tfrac{1}{n}J_n\right)Z$$

$$\Rightarrow \underbrace{Z^T Z}_{\sim \chi^2_n} = \underbrace{Z^T\left(I_n - \tfrac{1}{n}J_n\right)Z}_{\sum_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^2} + \underbrace{Z^T\left(\tfrac{1}{n}J_n\right)Z}_{n\left(\frac{\bar{X}-\mu}{\sigma}\right)}$$

$$\operatorname{rank}\left(I_n - \tfrac{1}{n}J_n\right) = n-1 \qquad \operatorname{rank}\left(\tfrac{1}{n}J_n\right) = 1$$

Cochran's theorem:

$$\underbrace{\sum_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{\frac{(n-1)S^2}{\sigma^2}} = Z^T\left(I_n - \tfrac{1}{n}J_n\right)Z \sim \chi^2_{n-1} \Rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 \sim \chi^2_1, \text{ and } \frac{(n-1)S^2}{\sigma^2} \perp\!\!\!\perp \left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right)^2$$

//

Global F tests for linear regression

Test for a population ~~mean~~ proportion

Suppose $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p)$. We want to test

$$H_0 : p = p_0 \qquad H_A : p \neq p_0$$

Wald test: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ or $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$

> Why is a t-test not appropriate?

- we don't have to separately estimate the variance (only estimating p)

Linear regression: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $\quad Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

Test for logistic regression

$$Y_i \sim Bernoulli(p_i) \qquad \log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i$$

We want to test

$$H_0 : C\beta = \gamma_0 \qquad H_A : C\beta \neq \gamma_0$$

$$(C\hat{\beta} - \gamma_0)^T (C \mathcal{I}^{-1}(\hat{\beta}) C^T)^{-1} (C\hat{\beta} - \gamma_0) \approx \chi^2_{\varepsilon} \text{ under } H_0$$

> Why is a t-test not appropriate?

- No separate variance term (σ^2) to estimate

Philosophical question

- If $X_1, \ldots, X_n$ are iid from a population with mean μ and variance σ^2, then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \xrightarrow{d} N(0,1)$
- If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \sim t_{n-1}$
- **Position 1:** For any reasonable sample size, the test statistic is approximately normal. And we never really have data from a normal distribution, so the t distribution is an approximation anyway. So always use the normal distribution
- **Position 2:** We always have a finite sample size, so our test statistic is never truly normal. And the t distribution is more conservative than the normal (heavier tails). So always use the t distribution

> With which position do you agree?