

# 18.650 – Fundamentals of Statistics

## 4. Hypothesis testing

# Goals

We have seen the basic notions of hypothesis testing:

- ▶ Hypotheses  $H_0/H_1$ , not symmetric, H0:status quo, H1: discovery
- ▶ Type 1/Type 2 error, level and power
- ▶ Test statistics and rejection region
- ▶ p-value

Our tests were based on CLT (and sometimes Slutsky)...

- ▶ What if data is Gaussian,  $\sigma^2$  is unknown and Slutsky does not apply? T-test 需要假设样本服从正态分布
- ▶ Can we use asymptotic normality of MLE? 不止有CLT可以有渐进正态，MLE也可以有渐进协方差矩阵(I^-1)。 如果感兴趣的不是均值，就不能用CLT了。 Wald's test
- ▶ Tests about multivariate parameters  $\theta = (\theta_1, \dots, \theta_d)$  (e.g.:  $\theta_1 = \theta_2$ )? Implicit hypotheses
- ▶ More complex tests: "Does my data follow a Gaussian distribution"? Goodness of fit.

# **Parametric hypothesis testing**

# Clinical trials

Let us go through an example to remind the main notions of hypothesis testing.

- ▶ Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- ▶ To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- ▶ We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

# Notation and modelling

- ▶ Let  $\Delta_d > 0$  denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- ▶ Let  $\Delta_c \geq 0$  denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.
- ▶ We want to know if  $\Delta_d > \Delta_c$
- ▶ We observe two independent samples:
  - ▶  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\Delta_d, \sigma_d^2)$  from the test group and
  - ▶  $Y_1, \dots, Y_m \stackrel{iid}{\sim} \mathcal{N}(\Delta_c, \sigma_c^2)$  from the control group.

# Hypothesis testing

- Hypotheses:

$$H_0 : \Delta_c = \Delta_d \quad \text{vs.} \quad H_1 : \Delta_d > \Delta_c$$

- Since the data is Gaussian by assumption we don't need the CLT

- We have

$$\bar{X}_n \sim \mathcal{N}\left(\Delta_d, \frac{\sigma_d^2}{n}\right) \quad \text{and} \quad \bar{Y}_m \sim \mathcal{N}\left(\Delta_c, \frac{\sigma_c^2}{m}\right)$$

- Therefore

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\sigma_d^2}{n} + \frac{\sigma_c^2}{m}}} \sim \mathcal{N}(0, 1)$$

方差可加

# Asymptotic test

- ▶ Assume that  $m = cn$  and  $n \rightarrow \infty$
- ▶ Using Slutsky's lemma, we also have

我们并不知道这个，但是在H0下，他们相等 (alpha水平是假设在H0下拒绝的概率)。

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\sigma_d^2}{n} + \frac{\sigma_c^2}{m}}} \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, 1)$$

m也要  
但是这里assume m is a constant times n

where

scaling by 1/(n-1) leads to an unbiased estimator

$$\hat{\sigma}_d^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad \text{and} \quad \hat{\sigma}_c^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y}_m)^2$$

- ▶ We get the the following test at asymptotic level  $\alpha$ :

$$R_\psi = \left\{ \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}} > q_\alpha \right\} \stackrel{=} {q_\alpha \text{ is } (1-\alpha) \text{ quantile of } \mathcal{N}(0, 1)}$$

- ▶ This is one-sided, two-sample test.

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\sigma_d^2}{n} + \frac{\sigma_c^2}{m}}} \xrightarrow{w} N(0, 1) \quad (\text{actually equal}).$$

$$\xrightarrow{n \rightarrow \infty} P \quad | \quad \text{by CLT}$$

→ Slutsky

# Asymptotic test

- ▶ Example  $n = 70, m = 50, \bar{X}_n = 156.4, \bar{Y}_m = 132.7, \hat{\sigma}_d^2 = 5198.4, \hat{\sigma}_c^2 = 3867.0,$

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{70} + \frac{3867.0}{50}}} = 1.57$$

Since  $q_{5\%} = 1.645$ , we *fail to reject  $H_0$*

- ▶ We can also compute the p-value:

$$\text{p-value} = \text{IP}(N(0,1) > 1.57) = 0.0582$$

p-value is the probability that i would reject even more than that i'm currently doing  
It's probability that the distribution of my test statistic is larger than the actual value of my test statistic

## PHARMACOLOGICAL DRUG TRIAL RESULTS



OUR TRIALS SHOW THAT  
THE NEW DRUG PERFORMS  
NO BETTER THAN PLACEBO

MAYBE WE SHOULD  
INVEST IN PLACEBOS

CHRIS  
MADDEN

# Small sample size

- ▶ What if  $n = 20, m = 12?$
- ▶ We cannot realistically apply Slutsky's lemma
- ▶ We needed it to find the (asymptotic) distribution of quantities of the form

如果你的数据是正态分布，那么 $\bar{X} - \mu$ 也是正态分布。相当于一个均值为0的正态分布。(linear combination)

$$\frac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}^2}}$$

样本小，所以用不了Slutsky  
CLT需要用Slutsky，但是我们并不需要它  
我们只需要用样本方差替代总体方差的时候，使用Slutsky

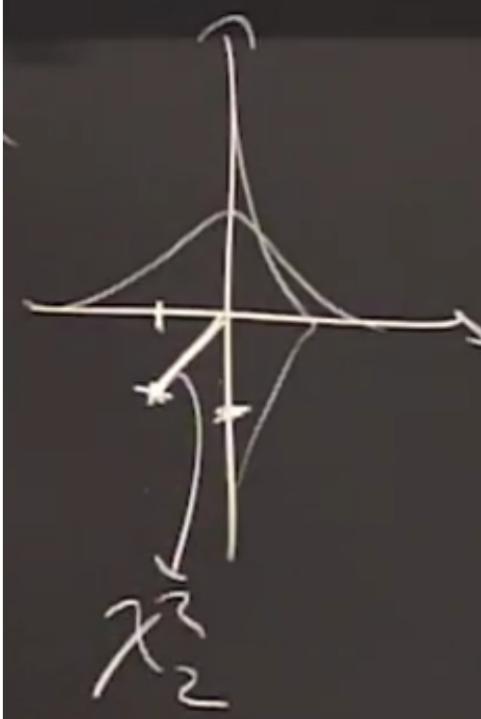
when  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ .

- ▶ It turns out that this distribution *does not depend on  $\mu$  or  $\sigma$*  so we can compute its *quantiles*

# The $\chi^2$ distribution

measuring the length of Gaussian vector(square length)

二维正态分布点离原点的距离的平方。



## Definition

For a positive integer  $d$ , the  $\chi^2$  (*pronounced “Kai-squared”*) *distribution with  $d$  degrees of freedom* is the law of the random variable  $Z_1^2 + Z_2^2 + \dots + Z_d^2$ , where  $Z_1, \dots, Z_d \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ .

这里的d是几个变量

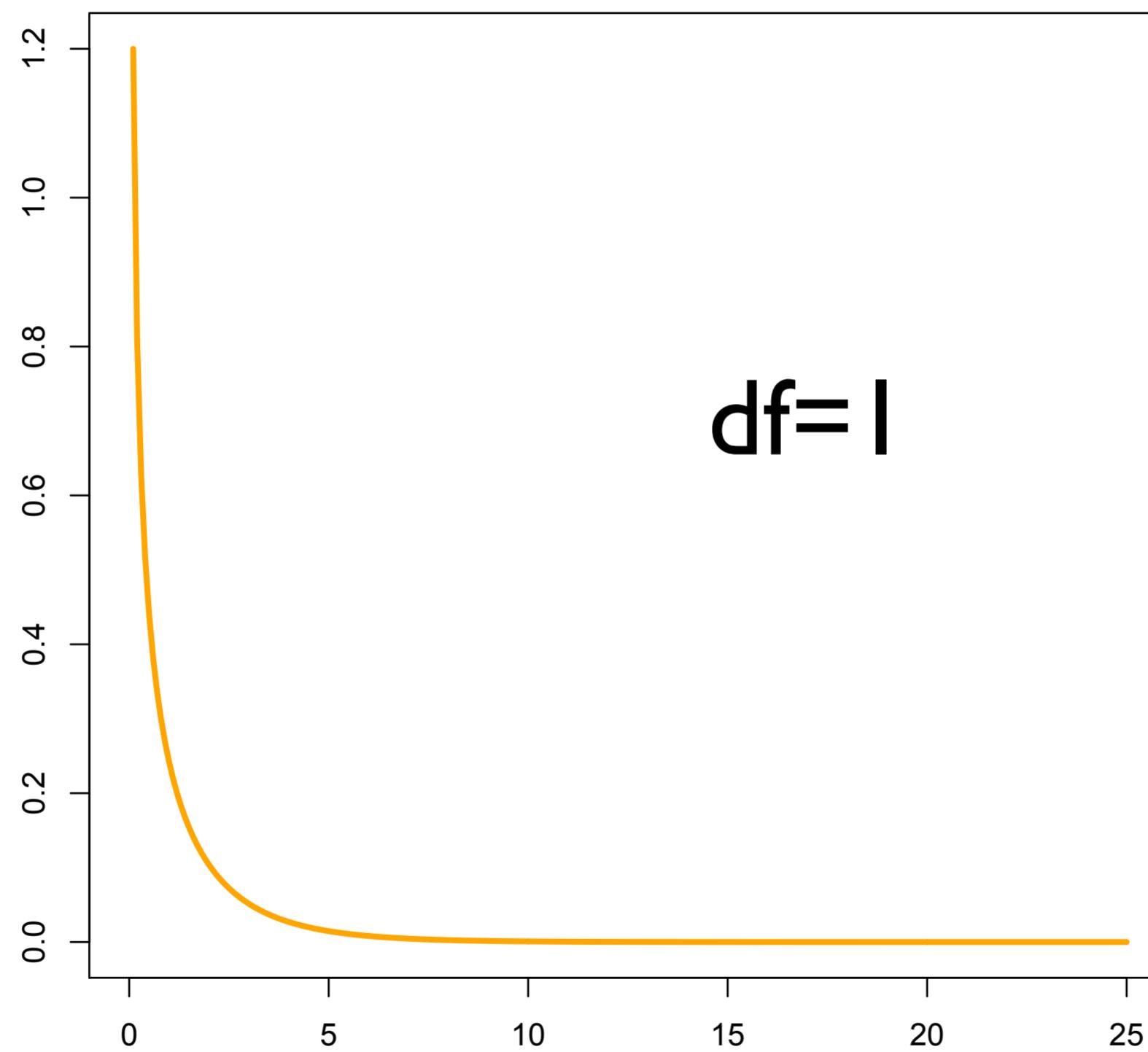
## Examples:

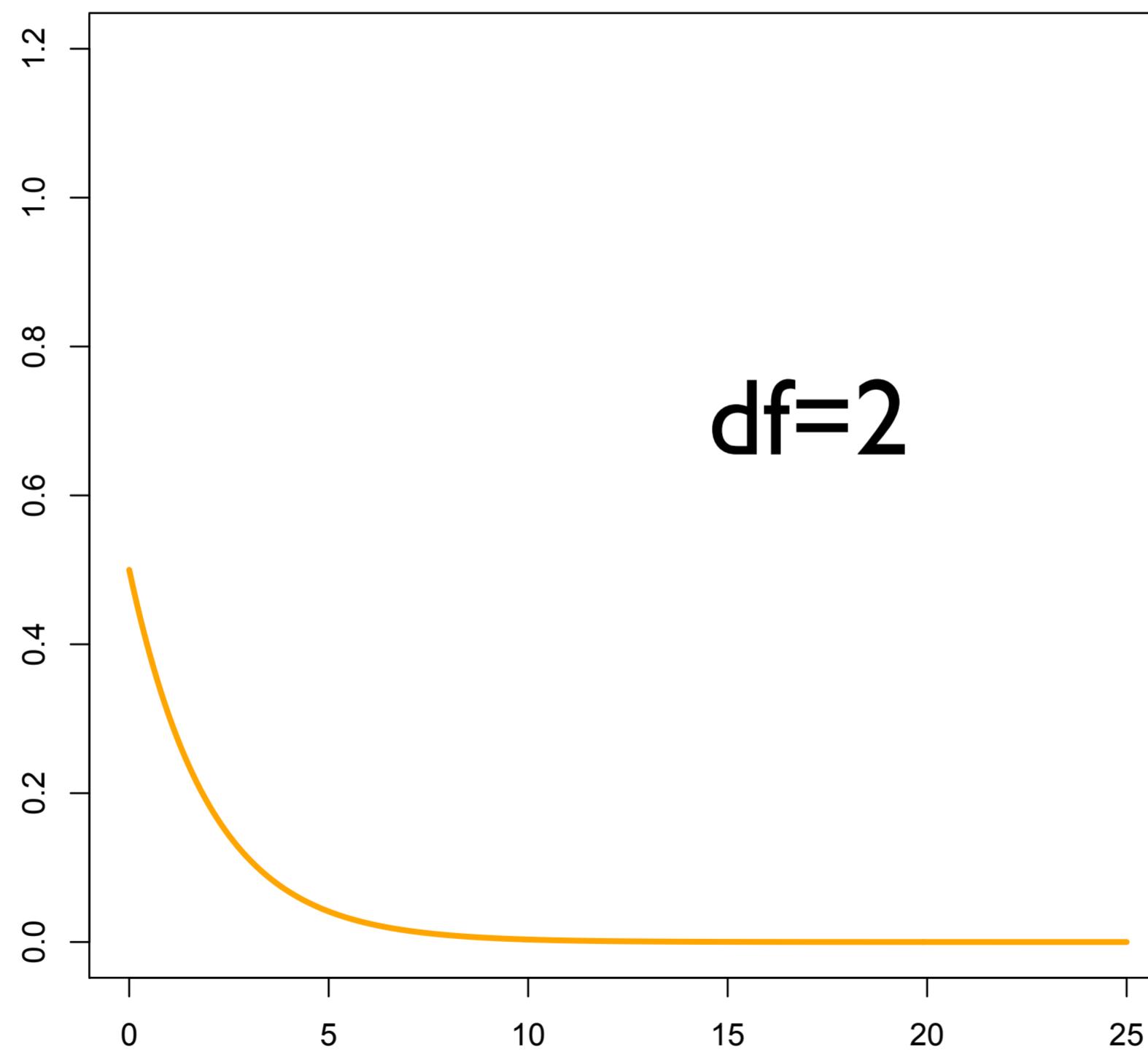
这里的d是维度

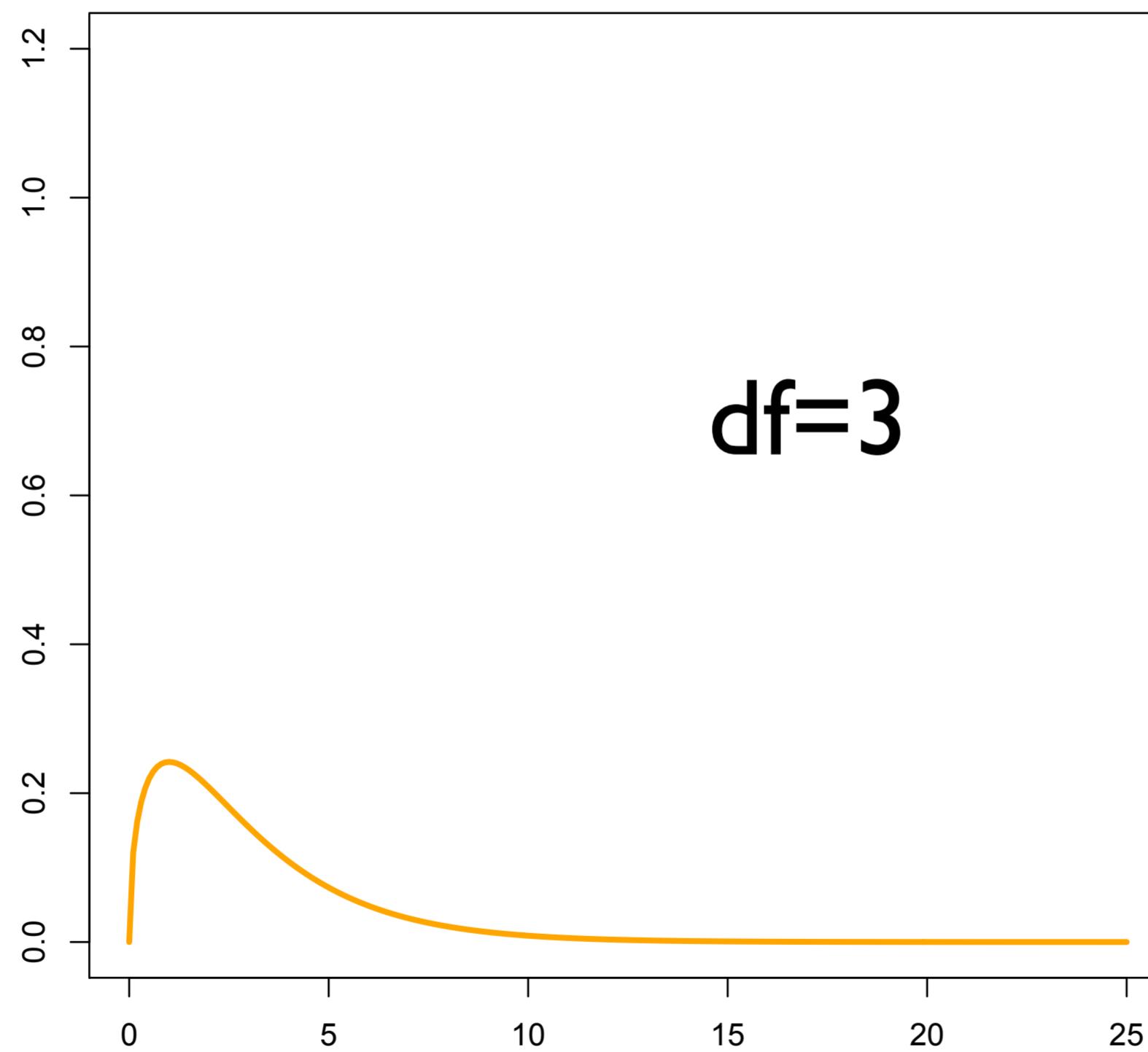
L2 norm  
Euclidean norm

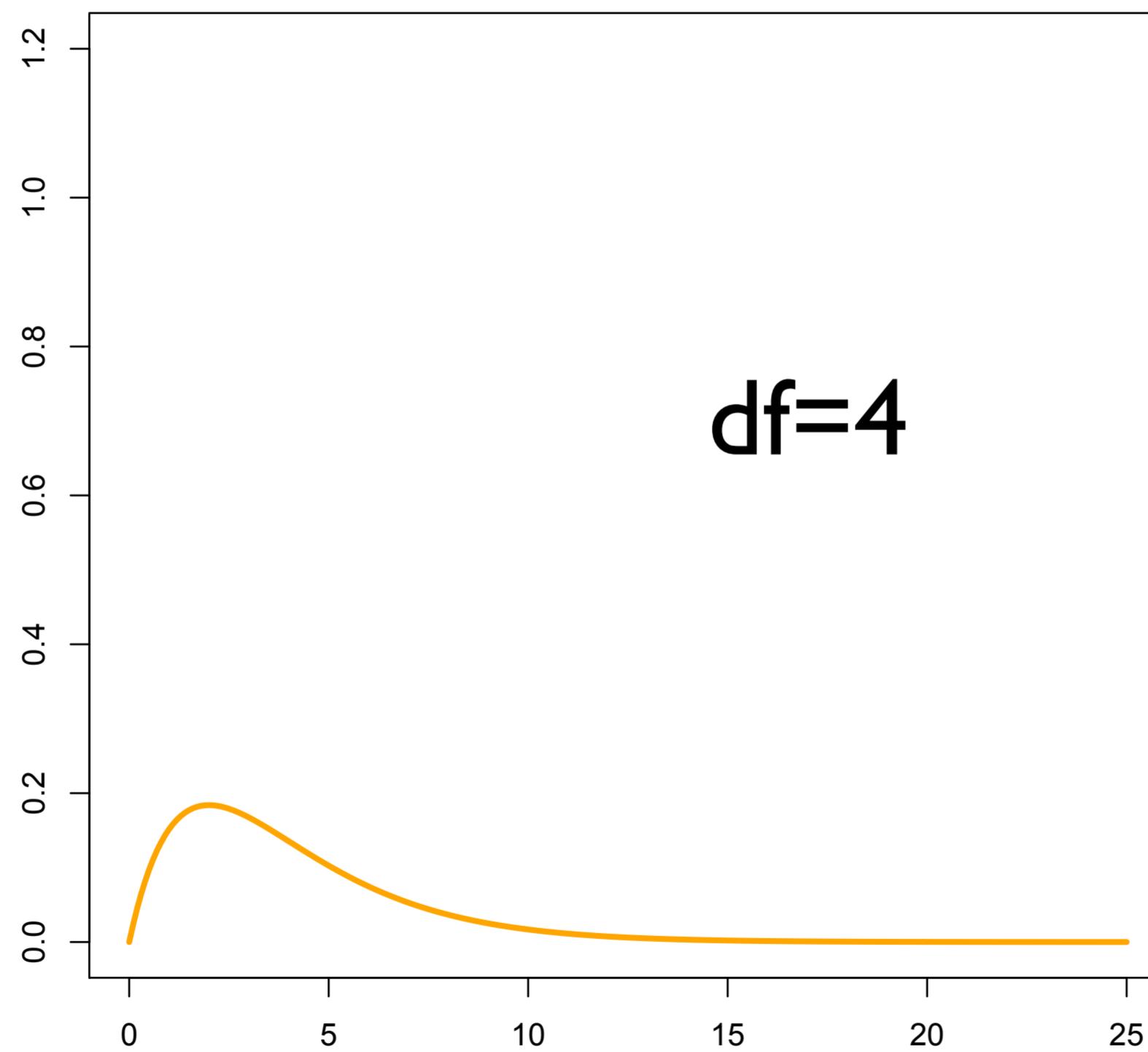
- If  $Z \sim \mathcal{N}_d(\mathbf{0}, I_d)$ , then  $\|Z\|_2^2 \sim \chi_d^2$
- $\chi_2^2 = \text{Exp}(1/2)$ .

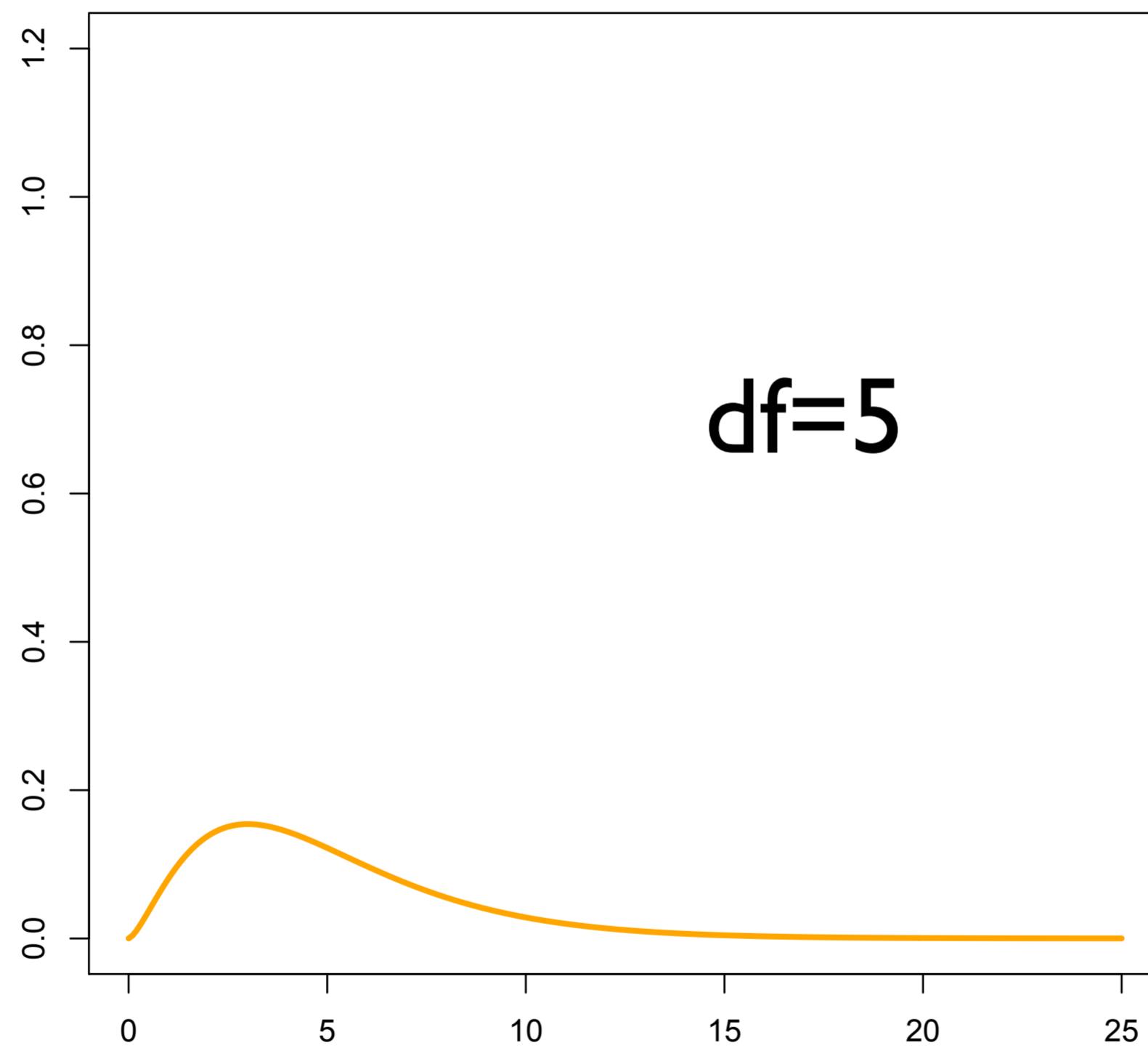
$$\text{pdf of } X \sim \chi_k^2 = \frac{1}{2^{k/2} \Gamma(\frac{k}{2})} x^{k/2 - 1} e^{-x/2}, x > 0$$

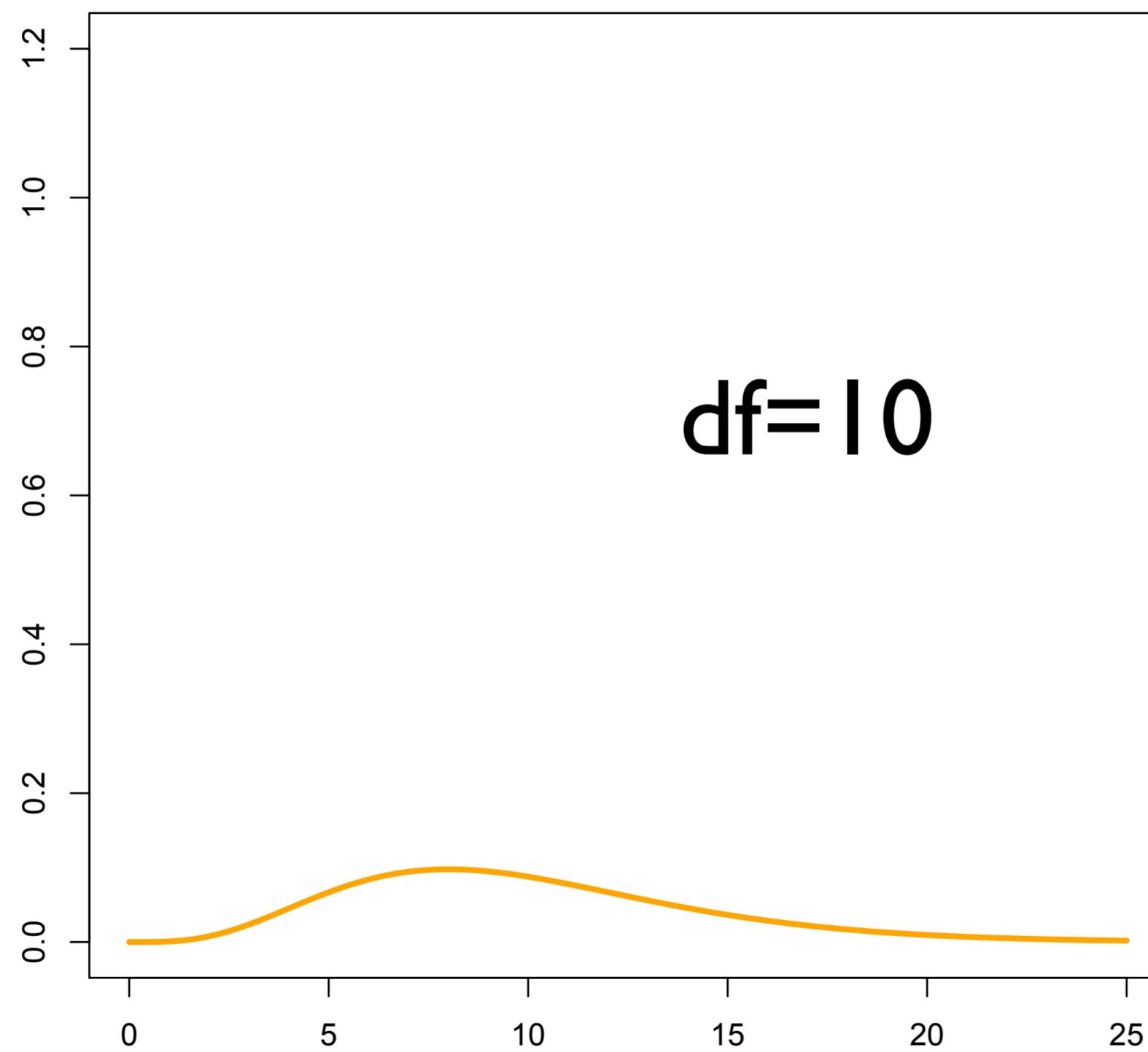


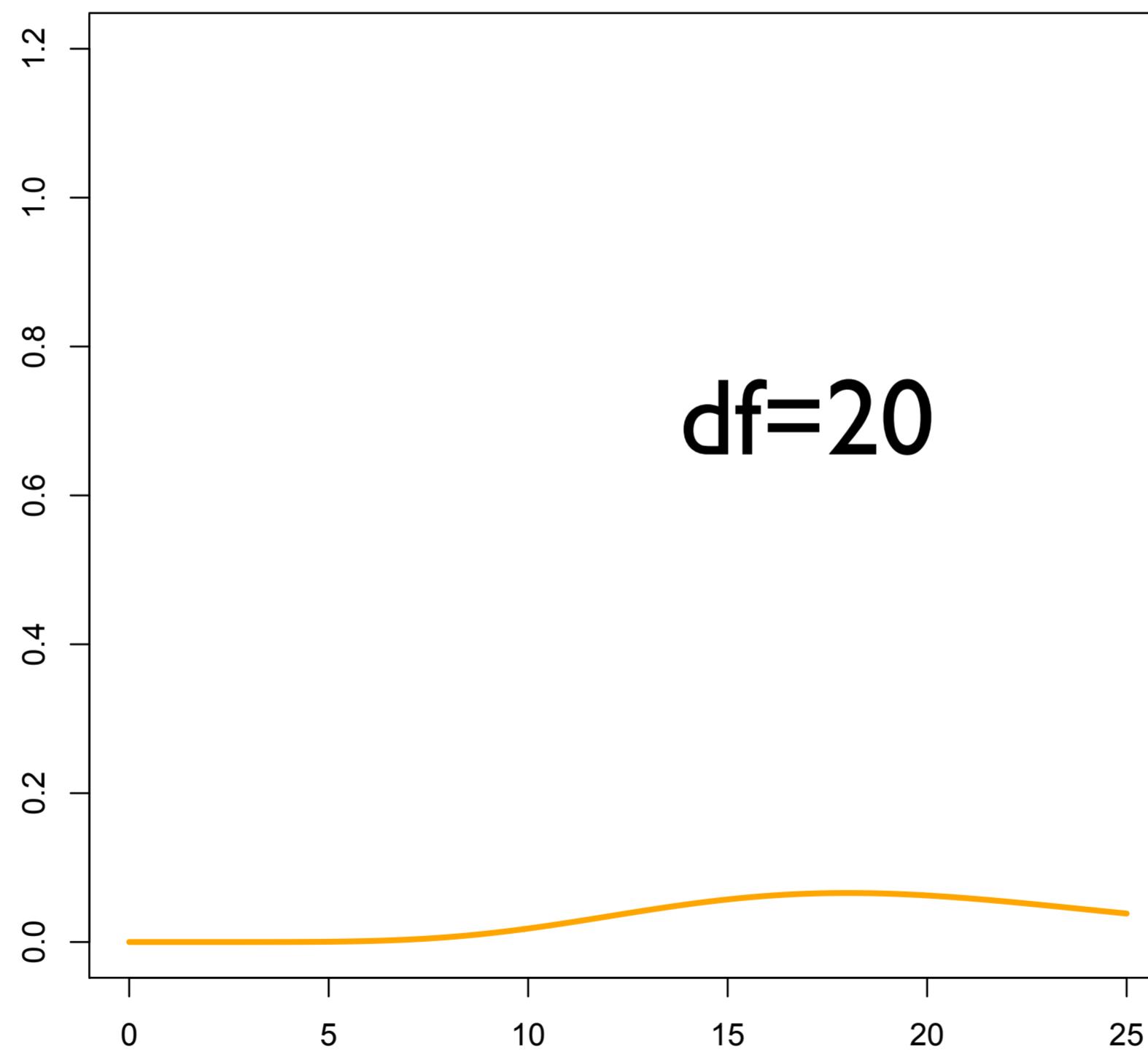












# Properties $\chi^2$ distribution (2)

## Definition

For a positive integer  $d$ , the  $\chi^2$  (*pronounced “Kai-squared”*) *distribution with  $d$  degrees of freedom* is the law of the random variable  $Z_1^2 + Z_2^2 + \dots + Z_d^2$ , where  $Z_1, \dots, Z_d \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ .

Properties: If  $V \sim \chi_k^2$ , then

- $\mathbb{E}[V] = \mathbb{E}[Z_1^2] + \dots + \mathbb{E}[Z_d^2] = d$
- $\text{var}[V] = \text{Var}[Z_1^2] + \dots + \text{Var}[Z_d^2] = 2d$

$$\text{Var}[Z_1^2] = \mathbb{E}[Z_1^4] - 1 = 3 - 1 = 2$$

third moment is measuring skewness  
forth moment is measuring kurtosis

如果d很大，那么 $\chi^2 \rightarrow N(d, 2d)$  CLT

# Important example: the sample variance

- Recall that the sample variance is given by

$$S_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$$

- Cochran's theorem states that for  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ , if  $S_n$  is the sample variance, then

- $\bar{X}_n \perp \!\!\! \perp S_n$ ; for all  $n$ .

independent ( $n-1$ )

- $$\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$$

这个是自由度， $n$ 个r.v.只有 $n-1$ 个自由度。他不是 $n$ 个独立的正态分布变量，有一个自由度丢失了。因为你知道，这些数加起来是0，但是当你把 $n$ 个独立的正态分布变量加起来的时候，你几乎肯定得不到0。

$$E[S_n] = \frac{n-1}{n} \sigma^2$$

- We often prefer the unbiased estimator of  $\sigma^2$ :

$$\tilde{S}_n = \frac{1}{n-1} \sum_{i=1}^{n-1} (X_i - \bar{X}_n)^2 = \frac{n}{n-1} S_n$$

$$E[\tilde{S}_n] = \frac{n}{n-1} E\left[\frac{\sigma^2}{n} \chi_{n-1}^2\right] = \frac{n\sigma^2}{n-1} \frac{n-1}{n} = \sigma^2$$

# Student's T distribution

## Definition

For a positive integer  $d$ , the *Student's T distribution with  $d$  degrees of freedom* (denoted by  $t_d$ ) is the law of the random variable  $\frac{Z}{\sqrt{V/d}}$ , where  $Z \sim \mathcal{N}(0, 1)$ ,  $V \sim \chi_d^2$  and  $Z \perp\!\!\!\perp V$  ( $Z$  is independent of  $V$ ).

这个是定义

这个的sample variance就是 V

# BIOMETRIKA.

---

## THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

### *Introduction.*

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large we may have precise information

# Who was Student?



This distribution was introduced by **William Sealy Gosset** (1876–1937) in 1908 while he worked for the Guinness brewery in Dublin, Ireland.

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$$
$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$$
$$\Rightarrow \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

if  $\bar{x}$  is Gaussian.

$$\Rightarrow \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sqrt{\sigma^2}} \sim t_{n-1}$$

两种t检验的理解方式  
 1 ,  $N(0,1)$ 除以自由度为n-1的卡方分布  
 2 ,  $x_{\bar{}}\text{是正态分布时}$ , 如果想用样本方差估计总体方差, 就要将标准正态分布换成自由度为n-1的t分布

# Student's $T$ test (one sample, two-sided)

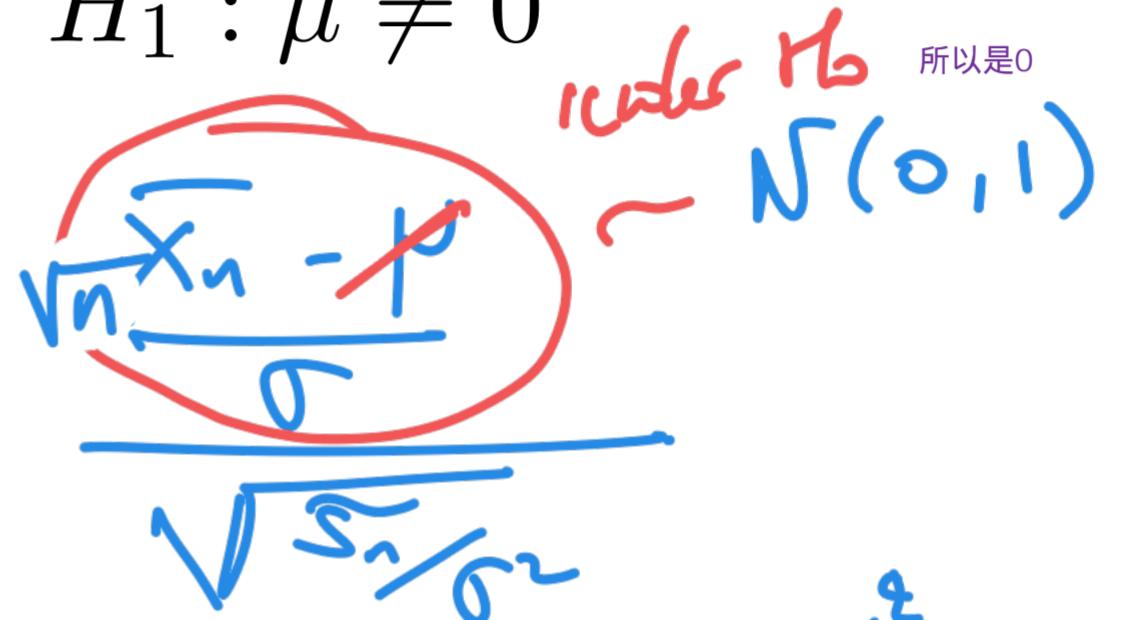
non-asymptotic: you don't know the distribution, have to assume.

- Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$  where both  $\mu$  and  $\sigma^2$  are unknown
- We want to test:

$$H_0 : \mu = 0, \quad \text{vs} \quad H_1 : \mu \neq 0$$

- Test statistic:

$$T_n = \frac{\bar{X}_n - \mu}{\sqrt{\tilde{S}_n/n}}$$



- Since  $\sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$  (under  $H_0$ ) and  $\tilde{S}_n / \sigma^2 \sim \chi_{n-1}^2$  are independent by Cochran's theorem, we have:

$$T_n \sim t_{n-1}$$

- Student's test with (non asymptotic) level  $\alpha \in (0, 1)$ :

$$\psi_\alpha = \mathbb{I}\{|T_n| > q_{\alpha/2}\},$$

where  $q_{\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of  $t_{n-1}$ .

# Student's T test (one sample, one-sided)

- We want to test:

等于mu0和小于mu0其实是一个意思，因为是无论mu取什么值，都小于mu0

$$H_0 : \mu \leq \mu_0, \quad \text{vs} \quad H_1 : \mu > \mu_0$$

- Test statistic:

$$T_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\tilde{S}_n}} \sim t_{n-1} \quad \text{under } H_0$$

under  $H_0$ .

- Student's test with (non asymptotic) level  $\alpha \in (0, 1)$ :

$$\psi_\alpha = \mathbb{I}\left\{ T_n > q_\alpha \right\},$$

where  $q_\alpha$  is the  $(1-\alpha)$ -quantile of  $t_{n-1}$

## Two-sample T-test

- ▶ Back to our cholesterol example. What happens for small sample sizes?
- ▶ We want to know the distribution of

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}}$$

- ▶ We have approximately

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}} \sim t_N$$

where

$$N = \frac{\left(\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}\right)^2}{\frac{\hat{\sigma}_d^4}{n^2(n-1)} + \frac{\hat{\sigma}_c^4}{m^2(m-1)}} \geq \min(n, m)$$

observation

N越大，越接近正态。  
N越小，越保守。

(Welch-Satterthwaite formula)

## Non-asymptotic test

- ▶ Example  $n = 70, m = 50, \bar{X}_n = 156.4, \bar{Y}_m = 132.7, \hat{\sigma}_d^2 = 5198.4, \hat{\sigma}_c^2 = 3867.0,$

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{70} + \frac{3867.0}{50}}} = \cancel{1.57} \quad 1.9248$$

- ▶ Using the shorthand formula  $N = \min(n, m) = 50$ , we get  $q_{5\%} = 1.68$  and

more conservative than normal  
tail has more weight

$$\text{p-value} = \text{P}[t_{50} > 1.57] = \cancel{0.0614} \quad 0.029974$$

- ▶ Using the W-S formula

$$N = \frac{\left(\frac{5198.4}{70} + \frac{3867.0}{50}\right)^2}{\frac{5198.4^2}{70^2(70-1)} + \frac{3867.0^2}{50^2(50-1)}} = 113.78$$

we round to 113.

- ▶ We get

$$\text{p-value} = \text{P}[t_{113} > 1.57] = \cancel{0.0596} \quad 0.2832$$

## Non-asymptotic test

- ▶ Example  $n = 20$ ,  $m = 12$ ,  $\bar{X}_n = 156.4$ ,  $\bar{Y}_m = 132.7$ ,  $\hat{\sigma}_d^2 = 5198.4$ ,  $\hat{\sigma}_c^2 = 3867.0$ ,

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{20} + \frac{3867.0}{12}}} = 0.982$$

- ▶ Using the shorthand formula  $N = \min(n, m) = 12$ , we get  $q_{5\%} = 1.73$  and

$$\text{p-value} = P[t_{12} > 0.982] = 17.27\%$$

- ▶ Using the W-S formula

$$N = \frac{\left(\frac{5198.4}{20} + \frac{3867.0}{12}\right)^2}{\frac{5198.4^2}{20^2(20-1)} + \frac{3867.0^2}{12^2(12-1)}} = 26.07$$

we round to 26.

- ▶ We get

$$\text{p-value} = P[t_{12} > 0.982] = 16.75\%$$

# Discussion

Advantage of Student's test: Non asymptotic / Can be run on small samples + *Can always use it for large sample sizes.*

large data + CLT 就可以使用

Drawback of Student's test: It relies on the assumption that the sample is Gaussian (soon we will see how to test this assumption)

# A test based on the MLE

- ▶ Consider an i.i.d. sample  $X_1, \dots, X_n$  with statistical model  $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , where  $\Theta \subseteq \mathbb{R}^d$  ( $d \geq 1$ ) and let  $\theta_0 \in \Theta$  be fixed and given.  $\theta^*$  is the true parameter
- ▶ Consider the following hypotheses:

$$\begin{cases} H_0 : \theta^* = \theta_0 \\ H_1 : \theta^* \neq \theta_0. \end{cases}$$

- ▶ Let  $\hat{\theta}^{MLE}$  be the MLE. Assume the MLE technical conditions are satisfied.
- ▶ If  $H_0$  is true, then

MLE converges to true parameter, so we can plug in the theta

$$\text{sample(under } H_0) \quad \sqrt{n} \frac{I(\hat{\theta}_n)}{I(\theta_0)}^{\frac{1}{2}} \times \left( \hat{\theta}_n^{MLE} - \theta_0 \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d(0, I_d)$$

true  $\sqrt{n} \frac{I(\theta^*)}{I(\theta_0)}^{\frac{1}{2}}$

MLE(under  $H_0$ )  $\sqrt{n} \frac{I(\hat{\theta}_n^{MLE})}{I(\theta_0)}^{\frac{1}{2}}$

## Wald's test

► Hence,

$$\|\hat{\theta} - \theta_0\|^2 \sim N_d(0, I_d)$$

$$\sum (\hat{\theta}_j - \theta_j)^2 \approx z_1^2 + z_0^2$$

$$n \|I(\hat{\theta})^{1/2} (\hat{\theta} - \theta_0)\|^2 \sim \underbrace{\|N_d(0, I_d)\|_F^2}_{\chi_d^2 \text{ if } H_0 \text{ is true}}$$

vector  $v \in \mathbb{R}^d$ ,  $\|v\|^2 = v^T v = \sum_{j=1}^d v_j^2$

$$(\hat{\theta} - \theta_0)^T (I(\hat{\theta})^{1/2})^T I(\hat{\theta})^{1/2} (\hat{\theta} - \theta_0)$$

$\underbrace{\|I(\hat{\theta})^{1/2}\|_F}_{\text{|| } I(\hat{\theta}) \text{ symmetric }}$

$$= (\hat{\theta} - \theta_0)^T I(\hat{\theta}) (\hat{\theta} - \theta_0)$$

$$\underbrace{n \left( \hat{\theta}_n^{MLE} - \theta_0 \right)^T I(\hat{\theta}^{MLE}) \left( \hat{\theta}_n^{MLE} - \theta_0 \right)}_{T_n} \xrightarrow[n \rightarrow \infty]{(d)} \chi_d^2$$

► Wald's test with asymptotic level  $\alpha \in (0, 1)$ :

$$\psi = \mathbb{I}\{T_n > q_\alpha\},$$

One dimension wald's test  
 $\sqrt{n} \frac{(\hat{\theta} - \theta_0)}{\sigma} \xrightarrow{(d)} N(0, 1)$

$\sigma^2$  known  
 reject if  $\frac{n}{\sigma^2} (\hat{\theta} - \theta_0)^2 > q_\alpha(\chi_1^2) \xrightarrow{n \rightarrow \infty} \chi_1^2$

$\Leftrightarrow \sqrt{n} \frac{|\hat{\theta} - \theta_0|}{\sigma} > \sqrt{q_\alpha(\chi_1^2)}$  inherently two-side

$$\sqrt{n} \frac{|\hat{\theta} - \theta_0|}{\sigma} > q_{\frac{\alpha}{2}}(N(0, 1))$$

where  $q_\alpha$  is the  $(1 - \alpha)$ -quantile of  $\chi_d^2$  (see tables).

► Remark: Wald's test is also **valid** if  $H_1$  has the form " $\theta > \theta_0$ " or " $\theta < \theta_0$ " or " $\theta = \theta_1$ "...

But less powerful

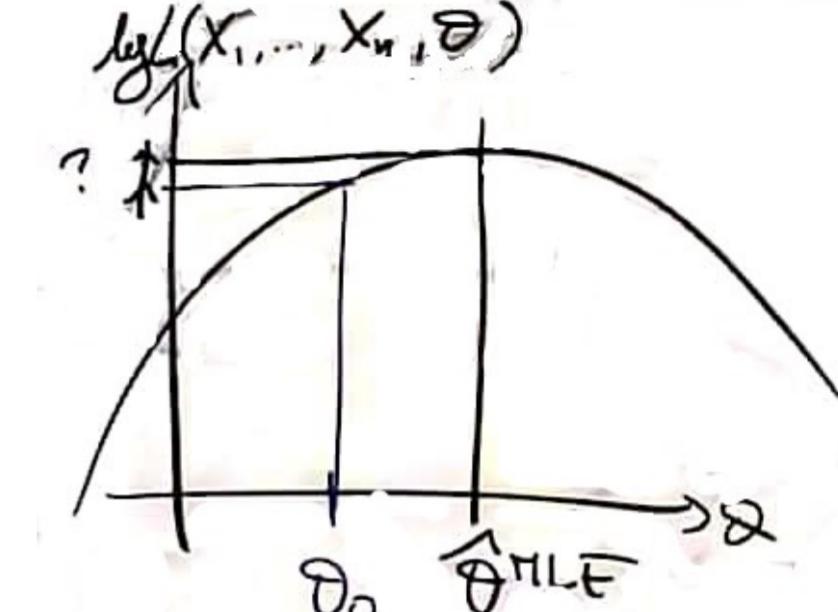
# A test based on the log-likelihood

- ▶ Consider an i.i.d. sample  $X_1, \dots, X_n$  with statistical model  $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , where  $\Theta \subseteq \mathbb{R}^d$  ( $d \geq 1$ ).
- ▶ Suppose the null hypothesis has the form

$$H_0 : (\theta_{r+1}, \dots, \theta_d) = (\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}),$$

a subset of the coordinates of theta

for some fixed and given numbers  $\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}$ .



- ▶ Let

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta) \quad (\text{MLE})$$

and

$$\hat{\theta}_n^c = \operatorname{argmax}_{\theta \in \Theta_0} \ell_n(\theta) \quad (\text{"constrained MLE"})$$

where  $\Theta_0 = \left\{ \theta \in \Theta : (\theta_{r+1}, \dots, \theta_d) = (\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}) \right\}$

# Likelihood ratio test

Test statistic:

$$T_n = 2 \left( \ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c) \right).$$

泰勒展开第二项的常数2

估计值

被限制的假设值

## Wilks' Theorem

Assume  $H_0$  is true and the MLE technical conditions are satisfied.

Then,

$$T_n \xrightarrow[n \rightarrow \infty]{(d)} \chi_{d-r}^2$$

d-r , constrained number of theta  
如果r是0，那么所有的theta都被限制了，就相当于问theta\_hat是不是等于theta\_0

Likelihood ratio test with asymptotic level  $\alpha \in (0, 1)$ :

$$\psi = \mathbb{I}\{T_n > q_\alpha\},$$

where  $q_\alpha$  is the  $(1 - \alpha)$ -quantile of  $\chi_{d-r}^2$  (see tables).

# Implicit hypotheses

- ▶ Let  $X_1, \dots, X_n$  be i.i.d. random variables and let  $\theta \in \mathbb{R}^d$  be a parameter associated with the distribution of  $X_1$  (e.g. a moment, the parameter of a statistical model, etc...)
- ▶ Let  $g : \mathbb{R}^d \xrightarrow{\text{function}} \mathbb{R}^k$  be continuously differentiable (with  $k < d$ ).
- ▶ Consider the following hypotheses:
$$\begin{cases} H_0 : & g(\theta) = 0 \\ H_1 : & g(\theta) \neq 0. \end{cases}$$
- ▶ E.g.  $g(\theta) = (\theta_1, \theta_2)$  ( $k = 2$ ), or  $g(\theta) = \theta_1 - \theta_2$  ( $k = 1$ ), or...

## Delta method

- ▶ Suppose an asymptotically normal estimator  $\hat{\theta}_n$  is available:

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d(0, \Sigma(\theta)).$$

- ▶ Delta method:

$$\sqrt{n} (g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_k(0, \Gamma(\theta)),$$

where  $\Gamma(\theta) = \nabla g(\theta)^\top \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k \times k}$ .

- ▶ Assume  $\Sigma(\theta)$  is invertible and  $\nabla g(\theta)$  has rank  $k$ . So,  $\Gamma(\theta)$  is invertible and

$$\sqrt{n} \Gamma(\theta)^{-1/2} (g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_k(0, \underline{\Gamma_k}).$$

# Wald's test for implicit hypotheses

- ▶ Then, by Slutsky's theorem, if  $\Gamma(\theta)$  is continuous in  $\theta$ ,

不知道theta是什么，我们只知道，under H0，我们只知道关于g(theta)的假设

$$\sqrt{n} \Gamma(\hat{\theta}_n)^{-1/2} (g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_k (0, I_k).$$

- ▶ Hence, if  $H_0$  is true, i.e.,  $g(\theta) = 0$ ,

$$\underbrace{ng(\hat{\theta}_n)^\top \Gamma^{-1}(\hat{\theta}_n) g(\hat{\theta}_n)}_{T_n} \xrightarrow[n \rightarrow \infty]{(d)} \chi_k^2.$$

- ▶ Test with asymptotic level  $\alpha$ :

$$\psi = \mathbb{I}\{\overline{T_n} > q_\alpha\},$$

where  $q_\alpha$  is the  $(1 - \alpha)$ -quantile of  $\chi_k^2$  (see tables).

# **Goodness of fit**

# Goodness of fit tests

Let  $X$  be a r.v. Given i.i.d copies of  $X$  we want to answer the following types of questions:

- ▶ Does  $X$  have distribution  $\mathcal{N}(0, 1)$ ? (Cf. Student's T distribution)
- ▶ Does  $X$  have distribution  $\mathcal{U}([0, 1])$ ?
- ▶ Does  $X$  have PMF  $p_1 = 0.3, p_2 = 0.5, p_3 = 0.2$

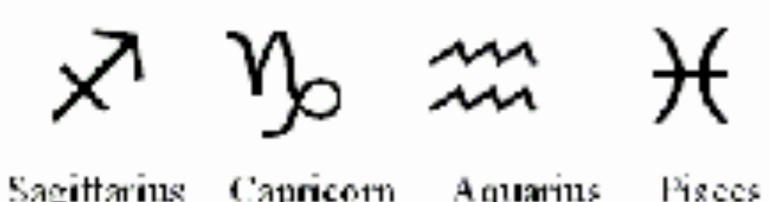
These are all *goodness of fit* (GoF) tests: we want to know if the hypothesized distribution is a good fit for the data.

Key characteristic of GoF tests: no parametric modeling.

# The zodiac sign of the most powerful people is....

Can your zodiac sign predict how successful you will be later in life?

Fortune magazine collected the signs of 256 heads of the Fortune 500.



Fyi:  
 $256/12 = 21.33$

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

# The zodiac sign of the most successful people is....

Sign	Count
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

In view of this data, is there statistical evidence that successful people are more likely to be born under some sign than others?

$$p^o = \left( \frac{1}{12}, \dots, \frac{1}{12} \right) \in \Delta_{12}$$

12 times

275 jurors with identified racial group.

We want to know if the jury is representative of the population of this county.

$$p^o = (0.72, 0.07, 0.12, 0.09)$$

$$K = 4$$

Race	White	Black	Hispanic	Other	Total
# jurors	205	26	25	19	275
proportion in county	0.72	0.07	0.12	0.09	1

# Discrete distribution

Let  $E = \{a_1, \dots, a_K\}$  be a finite space and  $(\mathbb{P}_p)_{p \in \Delta_K}$  be the family of all probability distributions on  $E$ :

$$p_j = \mathbb{P}[X = a_j]$$

- $\Delta_K = \left\{ p = (p_1, \dots, p_K) \in (0, 1)^K : \sum_{j=1}^K p_j = 1 \right\}.$
- For  $p \in \Delta_K$  and  $X \sim \mathbb{P}_p$ ,

$$\mathbb{P}_p[X = a_j] = p_j, \quad j = 1, \dots, K.$$

## Goodness of fit test

- ▶ Let  $X_1, \dots, X_n \stackrel{iid}{\sim} P_p$ , for some unknown  $p \in \Delta_K$ , and let  $p^0 \in \Delta_K$  be fixed.
- ▶ We want to test:

$$H_0: p = p^0 \text{ vs. } H_1: p \neq p^0$$

with asymptotic level  $\alpha \in (0, 1)$ .

- ▶ Example: If  $p^0 = (1/K, 1/K, \dots, 1/K)$ , we are testing whether  $P_p$  is *the uniform distribution* on  $E$ .

# Multinomial likelihood

$K = \# \text{ of modalities}$

- Likelihood of the model:

$$X \sim \text{Multinomial}(\underbrace{p_1, \dots, p_K}_{\vec{p}})$$

$$L_n(X_1, \dots, X_n, \mathbf{p}) = p_1^{N_1} p_2^{N_2} \cdots p_K^{N_K},$$

where  $N_j = \#\{i = 1, \dots, n : X_i = a_j\}$ .

- Let  $\hat{\mathbf{p}}$  be the MLE:

$$\hat{p}_j = \frac{N_j}{n}, \quad j = 1, \dots, K.$$



$\hat{\mathbf{p}}$  maximizes  $\log L_n(X_1, \dots, X_n, \mathbf{p})$  under the constraint

# $\chi^2$ test

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^0)^T \mathbf{1} = \sum_{i=1}^K (\hat{p}_i - p_i^0) = \sum_{i=1}^K \hat{p}_i - \sum_{i=1}^K p_i^0 = 0.$$

1, 这个乘以一个全是1的列向量应该是0  
2, 这个一定asymptotically normal

3, 但是  $N(0, I_K)$  和  $I_K$  的内积不是0 (因为 Gaussian 向量之间的和一定是一个 Gaussian r.v.)

- If  $H_0$  is true, then  $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^0)$  is asymptotically normal, and the following holds.

Theorem Under  $H_0$ :

4, 所以这个向量一定是一个与全是1的列向量垂直的Gaussian(内积为0)

5, 所以, 由于限制条件, 也就是所有p加起来要等于1。我们损失了一个自由度, 只有  $K-1$  维度了。  
(在这个全是1的列向量的方向上, 没有方差)

$$\underbrace{n \sum_{j=1}^K \frac{(\hat{\mathbf{p}}_j - \mathbf{p}_j^0)^2}{\mathbf{p}_j^0}}_{T_n \text{ Wald' test}} \xrightarrow[n \rightarrow \infty]{(d)} \chi_{K-1}^2.$$

support size - dimension of theta - 1

- $\chi^2$  test with asymptotic level  $\alpha$ :  $\psi_\alpha = \mathbb{I}\{T_n > q_\alpha\}$ , where  $q_\alpha$  is the  $(1 - \alpha)$ -quantile of  $\chi_{K-1}^2$ .

- Asymptotic  $p$ -value of this test:  $p$ -value =  $\mathbb{P}[Z > T_n | T_n]$ , where  $Z \sim \chi_{K-1}^2$  and  $Z \perp T_n$ .

# CDF and empirical CDF

Let  $X_1, \dots, X_n$  be i.i.d. real random variables. Recall the cdf of  $X_1$  is defined as:

$$F(t) = \mathbb{P}[X_1 \leq t], \quad \forall t \in \mathbb{R}.$$

**It completely characterizes the distribution of  $X_1$ .**

**Definition** step function

The *empirical cdf* of the sample  $X_1, \dots, X_n$  is defined as:

(a.k.a sample cdf)

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq t\} \\ &= \frac{\#\{i = 1, \dots, n : X_i \leq t\}}{n}, \quad \forall t \in \mathbb{R}. \end{aligned}$$

E[ $\mathbb{I}\{X < t\}$ ]  
random function(depend  
on random observations)  
count # of i which  $X_i < t$

# Consistency

By the LLN, for all  $t \in \mathbb{R}$ ,

$$F_n(t) \xrightarrow[n \rightarrow \infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (*Fundamental theorem of statistics*)

uniform convergence (the rate of convergence is uniform, irrespective of  $t$ )

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

If we had  $\bar{F}_n(t) \rightarrow \bar{F}(t)$   $\forall t$

$t$  if depends on  $n$  or other conditions, then it would not be uniform

$$\bar{F}_n(t) = F(t) + \frac{t}{n}$$

$$\sup_{t \in \mathbb{R}} |\underbrace{\bar{F}_n(t)}_{\frac{t}{n}} - \underbrace{F(t)}_{\frac{t}{n}}| = \infty$$

# Asymptotic normality

By the CLT, for all  $t \in \mathbb{R}$ ,

$$\sqrt{n} \left( F_n(t) - F(t) \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}\left(0, \frac{F(t)(1-F(t))}{n}\right).$$

## Donsker's Theorem

In the worst case, how far  $F_n$  can get to  $F$ .

Under the  $H_0$ , 也就是两个相等

If  $F$  is continuous, then

$$\sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \rightarrow \infty]{(d)} \sup_{0 \leq t \leq 1} |\mathbb{B}(t)|,$$

where  $\mathbb{B}$  is a Brownian bridge on  $[0, 1]$ .

在0和1时，方差都是0  
把在时间点0和1都在x轴上的  
随机运动叫Brownian bridge

Brownian motion: random process  
在0到t时间中布朗运动最大的方差  
a r.v.

# Goodness of fit for continuous distributions

- ▶ Let  $X_1, \dots, X_n$  be i.i.d. real random variables with unknown cdf  $F$  and let  $F^0$  be a **continuous** cdf.  
no point has mass
- ▶ Consider the two hypotheses:

$$H_0 : F = F^0 \quad \text{v.s.} \quad H_1 : F \neq F^0.$$

- ▶ Let  $F_n$  be the empirical cdf of the sample  $X_1, \dots, X_n$ .
- ▶ If  $F = F^0$ , then  $F_n(t) \approx F^0(t)$ , for all  $t \in [0, 1]$ .  $\mathbb{R}$

# Kolmogorov-Smirnov test

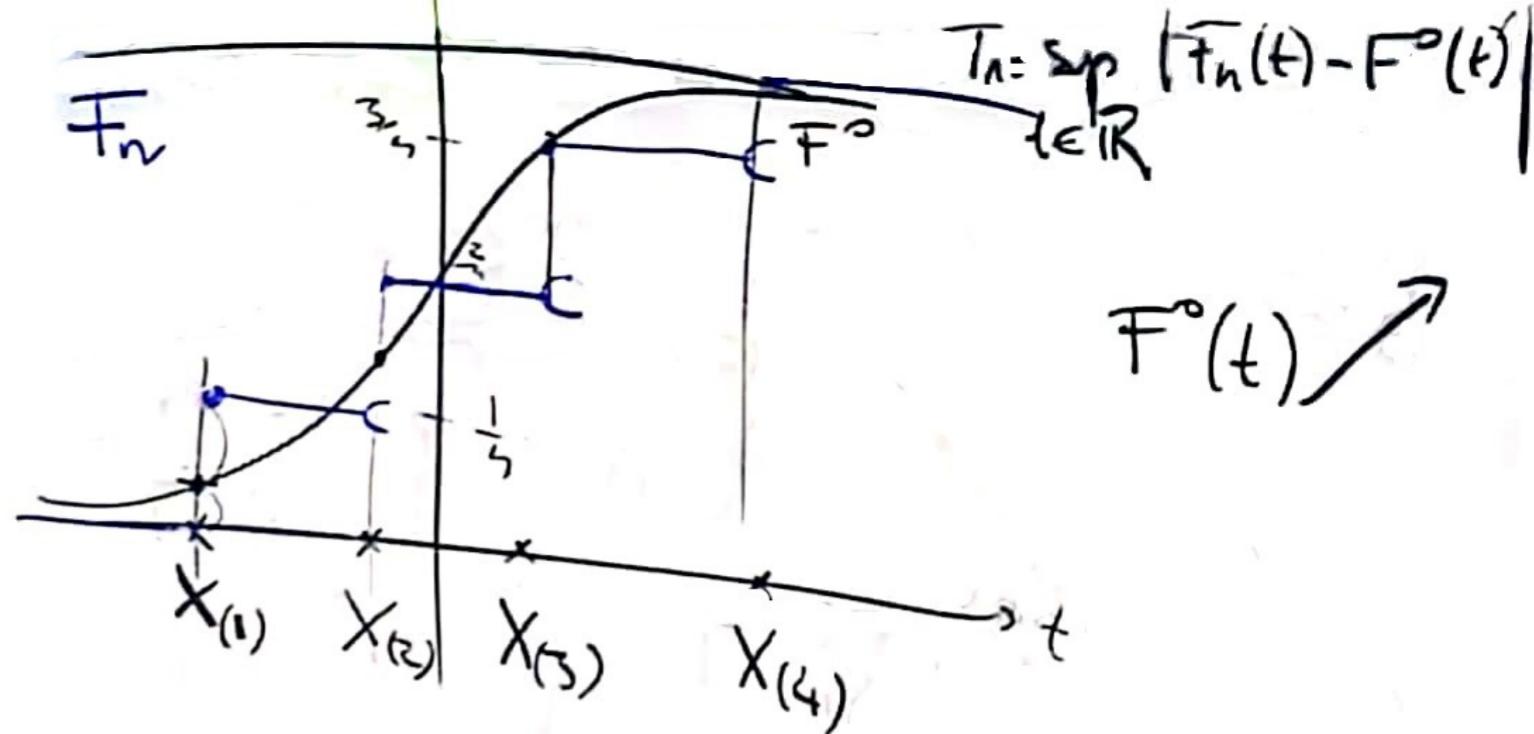
- Let  $T_n = \sup_{t \in \mathbb{R}} \sqrt{n} |F_n(t) - F^0(t)|$ .
- By Donsker's theorem, if  $H_0$  is true, then  $T_n \xrightarrow[n \rightarrow \infty]{(d)} Z$ , where  $Z$  has a known distribution (supremum of a Brownian bridge).
- **KS test with asymptotic level  $\alpha$ :**

$$\delta_\alpha^{KS} = \mathbb{I}\{T_n > q_\alpha\},$$

where  $q_\alpha$  is the  $(1 - \alpha)$ -quantile of  $Z$  (obtained in tables).

- p-value of KS test:  $\mathbb{P}[Z > T_n | T_n]$ .

# Computational issues



- ▶ In practice, how to compute  $T_n$  ?
- ▶  $F^0$  is non decreasing,  $F_n$  is **piecewise constant**, with jumps at  $t_i = X_i, i = 1, \dots, n$ .
- ▶ Let  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  be the reordered sample.
- ▶ The expression for  $T_n$  reduces to the following practical formula:

$$T_n = \sqrt{n} \max_{i=1,\dots,n} \left\{ \max \left( \left| \frac{i-1}{n} - F^0(X_{(i)}) \right|, \left| \frac{i}{n} - F^0(X_{(i)}) \right| \right) \right\}.$$

i是序数  
 某一段的左端点      小于第i个观测值的比率

如果  $X_i$  的 CDF 是  $F^0$  的话  
 那么  $F^0(X_i) \sim \text{Unif}(0,1)$

# Pivotal distribution

- $T_n$  is called a *pivotal statistic*: If  $H_0$  is true, the distribution of  $T_n$  does not depend on the distribution of the  $X_i$ 's and it is easy to reproduce it in simulations.
- Indeed, let  $U_i = F^0(X_i)$ ,  $i = 1, \dots, n$  and let  $G_n$  be the empirical cdf of  $U_1, \dots, U_n$ .
- If  $H_0$  is true, then  $U_1, \dots, U_n \stackrel{i.i.d.}{\sim} \text{Unif}([0, 1])$

$$\text{and } T_n = \sup_{0 \leq x \leq 1} \sqrt{n} |G_n(x) - x|.$$

↑  
empirical CDF of iid  
uniform r.v.

r.v.  
 $X \sim F, U \sim \text{Unif}[0, 1]$   
 $X = F^{-1}(U)$  (pseudo)  
 $P(X \leq t) = P(F^{-1}(U) \leq t)$   
 $= P(U \leq F(t))$  ( $F$  is increasing)  
 $= F(t)$

$F(x) \sim \text{Unif}[0, 1]$

# Quantiles and p-values

- ▶ For some large integer  $M$ :
  - ▶ Simulate  $M$  i.i.d. copies  $T_n^1, \dots, T_n^M$  of  $T_n$ ;
  - ▶ Estimate the  $(1 - \alpha)$ -quantile  $q_\alpha^{(n)}$  of  $T_n$  by taking the sample  $(1 - \alpha)$ -quantile  $\hat{q}_\alpha^{(n,M)}$  of  $T_n^1, \dots, T_n^M$ .

- ▶ Test with approximate level  $\alpha$ :

$$\delta_\alpha = \mathbb{I}\{T_n > \hat{q}_\alpha^{(n,M)}\}.$$

- ▶ Approximate p-value of this test:

$$\text{p-value} \approx \frac{\#\{j = 1, \dots, M : T_n^j > T_n\}}{M}.$$

# K-S table

## Kolmogorov–Smirnov Tables

---

Critical values,  $d_{alpha};(n)^a$ , of the maximum absolute difference between sample  $F_n(x)$  and population  $F(x)$  cumulative distribution.

---

Number of trials, $n$	Level of significance, $\alpha$			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893

# Other goodness of fit tests

We want to measure the distance between two functions:  $F_n(t)$  and  $F(t)$ . There are other ways, leading to other tests:

- Kolmogorov-Smirnov: worst possible deviation

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \quad L_\infty$$

- Cramér-Von Mises:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 dF(t) \quad L_2$$

$\stackrel{\text{:=}}{=} \mathbb{E}_{X \sim F} [\bar{F}_n(X) - \bar{F}(X)]^2$  pivotal

- Anderson-Darling:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t)$$

# Composite goodness of fit tests

What if I want to test: "Does  $X$  have Gaussian distribution?" but I don't know the parameters?  
Simple idea: plug-in

一般H0是deterministic，但是这里depends on estimated parameter.

$$\sup_{t \in \mathbb{R}} |F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t)|$$

where 此时，我们在选择H0的过程，就让H0与数据更相似。

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = S_n^2$$

and  $\Phi_{\hat{\mu}, \hat{\sigma}^2}(t)$  is the cdf of  $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ .

In this case **Donsker's theorem is no longer valid**. This is a common and serious mistake!

# Kolmogorov-Lilliefors test (1)

Instead, we compute the quantiles for the test statistic:

$$\sup_{t \in \mathbb{R}} |F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t)|$$

把数据用我们想要对比的正态分布的参数，进行normalize

They do not depend on unknown parameters!

只是表变了，值变小了，更容易拒绝H0了

This is the Kolmogorov-Lilliefors test.

The KL quantile could be  $x$  and the KS quantile  $y$  could be  $y > x$ . If your test statistic is  $c$  that is  $x < c < y$ , then KL will reject but KS will fail to reject.

# K-L table

Sample Size <i>N</i>	Level of Significance for $D = \text{Max}  F^*(X) - S_N(X) $				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231

LILIEFORS



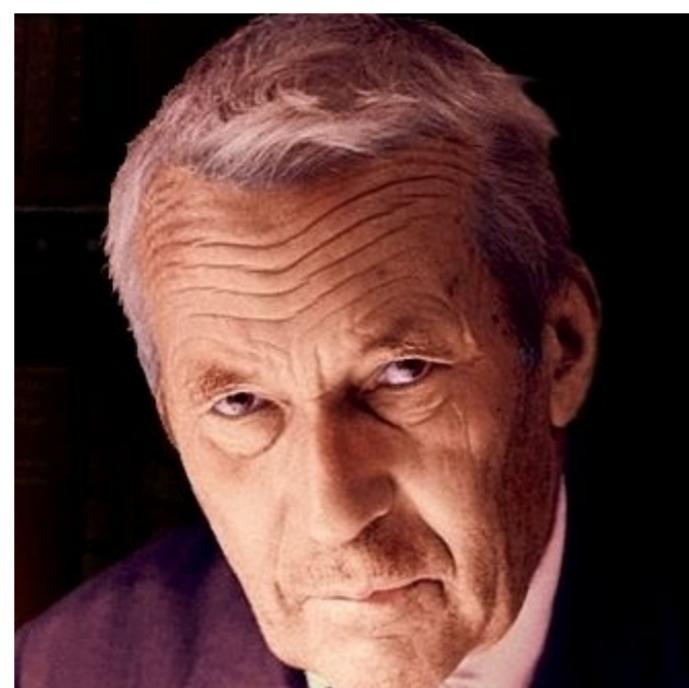
VON RISES



WALD



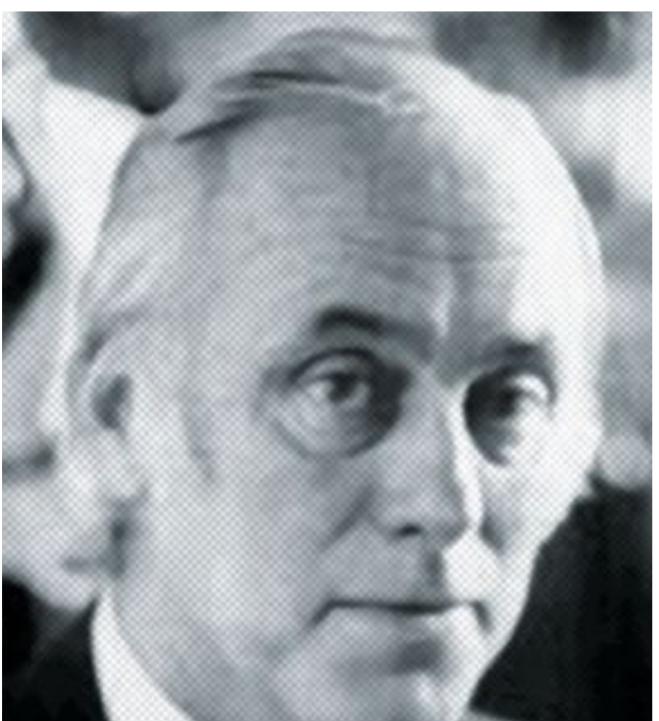
KOLLOKOBV



SIRNOV



WILKS



ANDERSON



DARLING



CRAMER

# Quantile-Quantile (QQ) plots (1)

- ▶ Provide a visual way to perform GoF tests
- ▶ Not formal test but quick and easy check to see if a distribution is plausible.
- ▶ Main idea: we want to check visually if the plot of  $F_n$  is close to that of  $F$  or equivalently if the plot of  $F_n^{-1}$  is close to that of  $F^{-1}$ .
- ▶ More convenient to check if the points

plot的是值而不是分数，相同quantile的值，一一比较

$$\left( F^{-1}\left(\frac{1}{n}\right), F_n^{-1}\left(\frac{1}{n}\right) \right), \left( F^{-1}\left(\frac{2}{n}\right), F_n^{-1}\left(\frac{2}{n}\right) \right), \dots, \left( F^{-1}\left(\frac{n-1}{n}\right), F_n^{-1}\left(\frac{n-1}{n}\right) \right)$$

$x_{(1)}$                                      $x_{(2)}$

are near the line  $y = x$ .

- ▶  $F_n$  is not technically invertible but we define

$$F_n^{-1}(i/n) = X_{(i)},$$

the  $i$ th largest observation.

IF  $F^{-1}$  is CDF of  $N(0, 1)$

$F^{-1}\left(\frac{i}{n}\right)$  is  $x_i$  such that

$$\int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{i}{n}$$

$$x_i = q_{1-\frac{i}{n}}$$

# Quantile-Quantile (QQ) plots (2)

y: observation , 数据 !

x: cdf (一般正态分布比 , 就是正态的 )

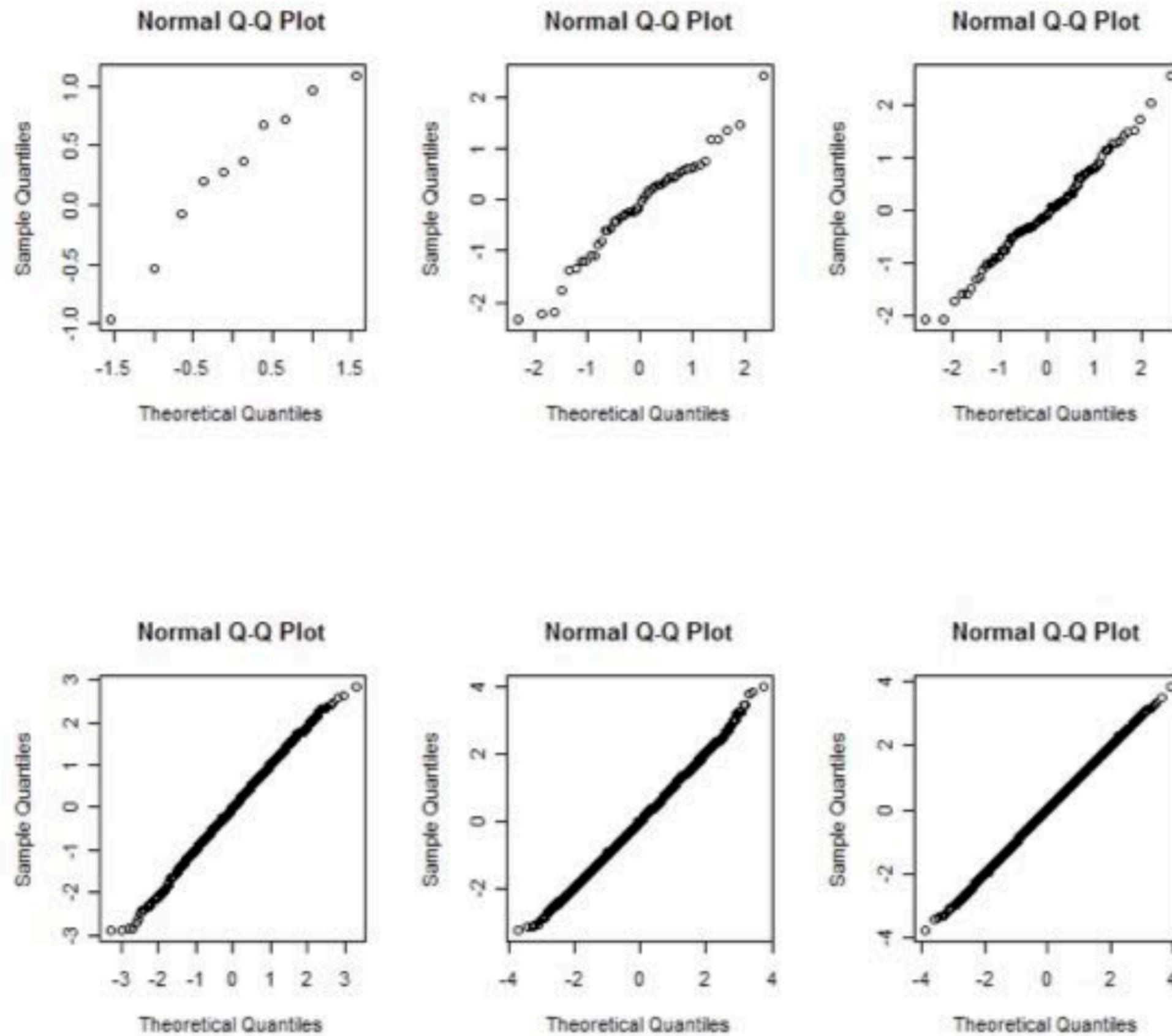


Figure 1: QQ-plots for samples of sizes 10, 50, 100, 1000, 5000, 10000 from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

# Quantile-Quantile (QQ) plots (3)

重尾更多的质量在尾部，同样的quantile， $T_n$ 就会更大  
 轻尾更多的质量在中间，同样的quantile， $T_n$ 就会更小

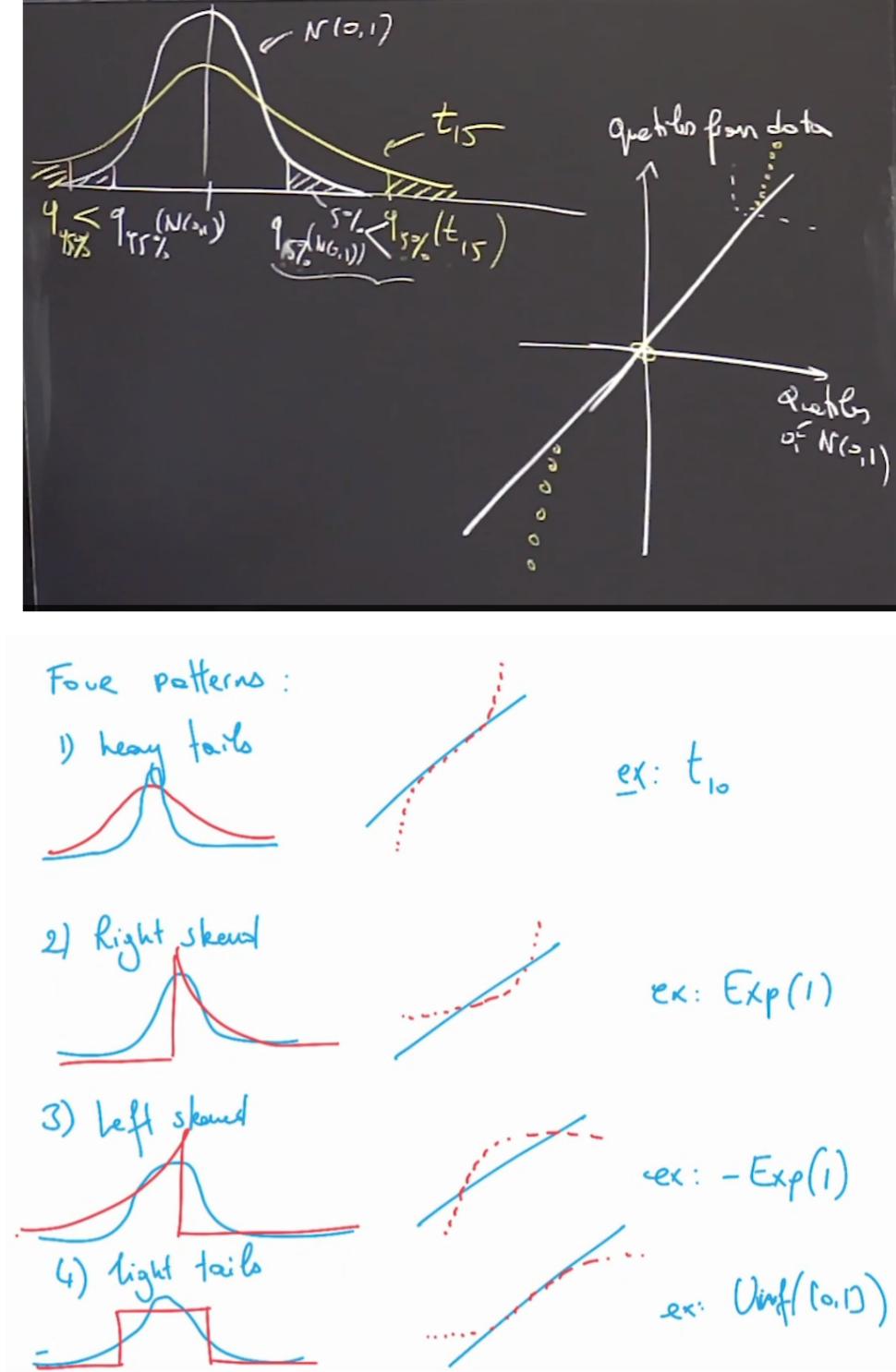
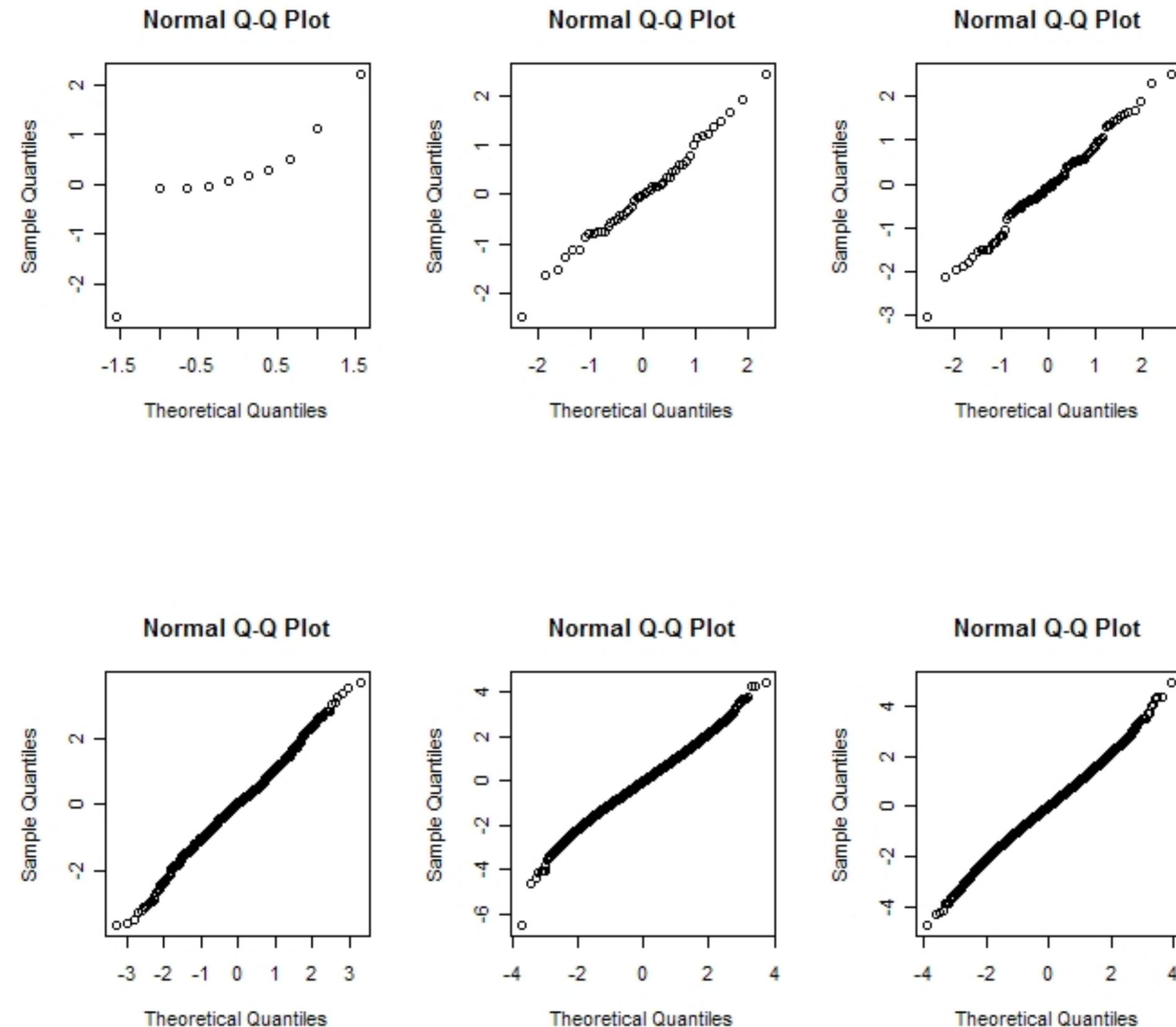


Figure 2: QQ-plots for samples of sizes 10, 50, 100, 1000, 5000, 10000 from a  $t_{15}$  distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.