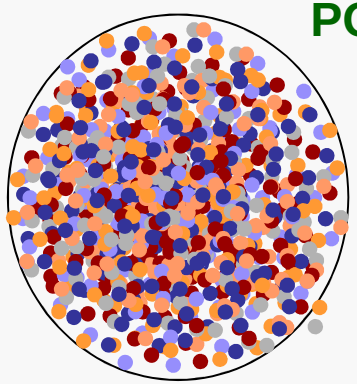
Lecture nº 3.

Estimation theory.

Echantillonnage et estimation.

Rappelons:



POPULATION P
(paramètres μ, σ, p)

Théorie d'échantillonnage

Comment choisir un échantillon représentatif de la populations

Théorie d'estimation

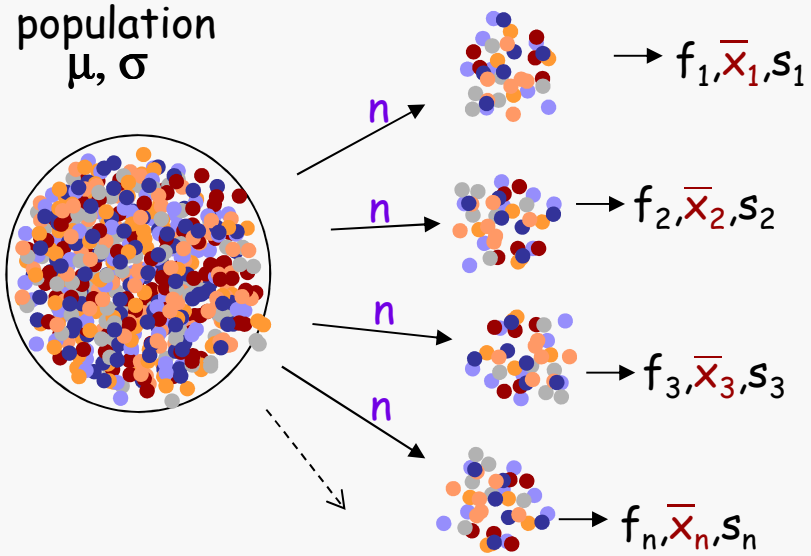
Comment estimer les paramètres d'une population à partir des statistiques recueillis sur son échantillon

ECHANTILLON E
(statistiques $\overline{x}$, S, f)

Ce cours est totalement consacré à la théorie d'éstimation.

Statistiques d'un échantillon.

On s'intéresse à l'étude d'un caractère X dans une population qu'on ne peut pas (on ne veut pas) recenser.



on extrait plusieurs échantillons :

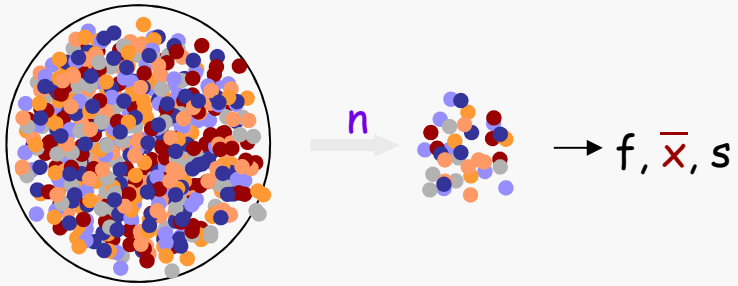
- représentatifs,
- aléatoires,
- de taille **n** fixée.

A PRIORI,
les caractéristiques des échantillons ne sont pas constantes;
elles varient (**fluctuent**) d'un échantillon à l'autre
(on dit qu'on observe des **fluctuations d'échantillonnage**).

DONC, à partir d'un échantillon on ne peut pas déterminer, mais
mais seulement **ESTIMER** les paramètres d'une population.

Estimateur biaisé ou non-biaisé?

Dans une population **P** le caractère **X** est décrit par des paramètres μ et σ.
Dans un échantillon de taille **n** le caractère **X** apparaît avec des statistiques $\bar{x}$ et **s**.



On estime les paramètres de la population **P** à partir de statistiques de l'échantillon.

On appelle une statistique un estimateur non-biaisé d'un paramètre si la valeur estimée de cette statistique est égale à la valeur du paramètre.

$$\mu = E(\bar{x})$$
$$\sigma^2 = E(S^2)$$

estimateurs non-biaisés

D'habitude $\sigma^2 \neq E(S^2)$

$$\sigma^2 = \frac{n}{n-1} E(S^2)$$

estimateur biaisé

Estimateur efficace.

Dans une population **P** le caractère **X** est décrit par des paramètres μ et σ.
Pour estimer la valeur de μ, on extrait un échantillon de cette population
et on calcule les trois statistiques décrivant la tendance centrale :

- moyenne M,
- médiane Me,
- modale Mo.

Chacune de ces statistiques est un estimateur non-biaisé de la moyenne:

$$E(M) = E(Me) = E(Mo) = \mu$$

QUESTION : Laquelle approche le mieux la moyenne μ de la population?

Parmi différentes statistiques du même paramètre on appelle meilleur estimateur (estimateur efficace) celle dont la variance est moindre..

Dans l'analyse statistique on veux toujours avoir
des estimateurs non-biaisés et efficaces des propriétés des populations,
mais ce n'est pas toujours facile (parfois impossible).

Estimation ponctuelle.

En pratique, on dispose souvent d'un seul échantillon de la population, de taille n .
La meilleure estimation ponctuelle *(par un seul nombre)*
des paramètres de la population est la suivante:

- la moyenne μ de la population est égale à la moyenne $\bar{x}$ de l'échantillon ;

$$\mu = \bar{x}$$

- la variance σ^2 de la population est égale à la variance estimée S_e^2 :

$$\sigma^2 = S_e^2 = \frac{n}{n-1} S^2$$

- la probabilité **p** d'apparition d'un caractère **C** dans la population est égale à sa fréquence d'apparition **f** dans l'échantillon.

$$p = f$$

A cause de fluctuations d'échantillonnage
estimation ponctuelle a un sens si et seulement si elle est accompagnée d'**évaluation de l'incertitude de l'estimation**.

Incertitude de l'estimation.

Incertitude de l'estimation est défini comme la valeur absolue de différence entre le paramètre de la population et son estimateur:

$$\Delta\mu = |\mu - \bar{x}|$$
$$\Delta\sigma^2 = |\sigma^2 - S^2|$$
$$\Delta p = |p - f|$$

Loi empirique:
pour les grands échantillons (n >30),
95.45 % de fois l'incertitude de l'estimation est plus petite que $2\sigma_x / \sqrt{n}$ (x = μ, σ ou p).

EXEMPLE 1: Un concessionnaire veut savoir combien de temps une voiture neuve qu'il vend peut rouler sans réparations. Un échantillon de 250 voitures a donné un temps moyen de 2.6 années avec l'écart type de 1.2 ans.

$$\mu = \bar{x} = 2.6\ ans$$ ← estimation ponctuelle

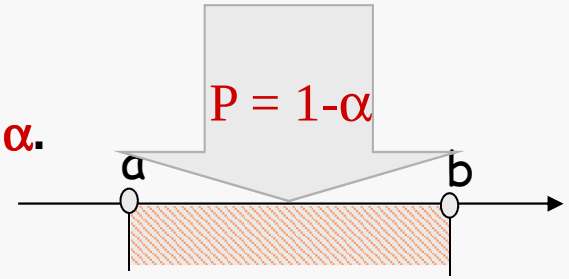
$$\Delta\mu = \frac{2 \cdot 1.2}{\sqrt{250}} = 0.152\ ans$$ ← incertitude de l'estimation

Estimation par intervalle de confiance.

PRINCIPE

- Soit : V – valeur d'un paramètre de la population P;
  E – valeur de la statistique correspondante
  dans un échantillon de taille n.

  - on choisi un nombre $\alpha \in (0, 1)$;
  - on détermine l'intervalle]a, b[tel, que la probabilité de se tromper en affirmant que $V \in (a, b)$ est égale α.



$$P\left(V \notin]a, b[\right) = \alpha$$

ATTENTION :

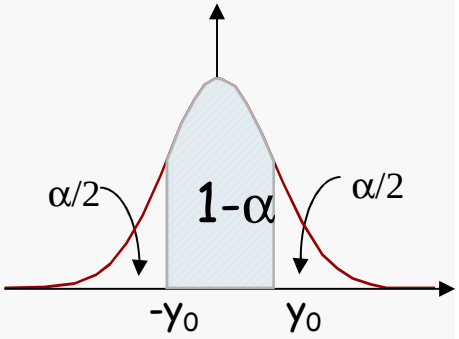
La tradition affecte aux réels α
prioritairement les valeurs 0.05 et 0.01.
Il faut donc lire α comme 'une faible probabilité'.

- En pratique, pour construire l'intervalle de confiance d'un paramètre, il faut connaître sa distribution de probabilité.

Estimation par un intervalle de confiance.

- Grands échantillons (n > 30):
  - quelle que soit la loi réellement suivie par la variable étudiée **X**, on peut toujours l'approximer par la loi normale.
  - Si **X** → N(**μ**, **σ**), alors **Y = (x-μ)/σ** → N(0, 1).
  - Pour un niveau de risque α choisi les tables statistiques donnent la valeur y_0 telle, que

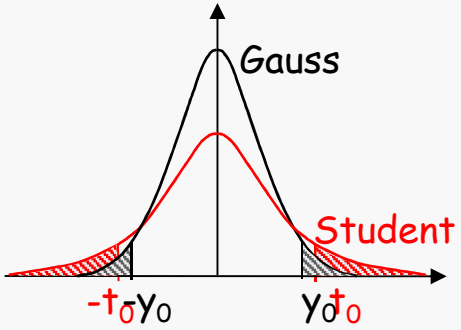
$$P(-y_0 < Y < y_0) = 1 - \alpha$$



]-y_0, y_0[- intervalle de confiance de variable **y**.
Connaissant y_0, on peut remonter à l'intervalle de confiance de **x**.

- Petits échantillons (n < 30) issus de la population normale.
  - Si **x** → N(**μ**, **σ**), la variable **t = (x-μ)/σ** suit la loi de Student à (n-1) degrés de liberté.
  - Les tables donnent la valeur t_0 telle, que

$$P(-t_o < t < t_0) = 1 - \alpha$$



ATTENTION:
Pour le même risque **α,**
l'intervalle de confiance de Student
est toujours plus large celui de la loi normale.

Intervalle de confiance d'un pourcentage.

On étudie l'apparition du caractère **C** chez les individus de la population **P**.

Soit:

nf - le nombre d'individus présentant ce caractère dans un échantillon de taille **n**.

nf suit la distribution binomiale B(n, p) (on possède, ou non, le caractère C).

QUESTION:

f est un estimateur ponctuel de p.

Au niveau de confiance α, dans quelles limites est contenue la vraie valeur de p ?

Si : • **n** est grand
• **p** n'est voisin ni de 0, ni de 1

$\Rightarrow$ **B(n, p)** $\rightarrow$ N (np, npq)

(*théorème central limite*)

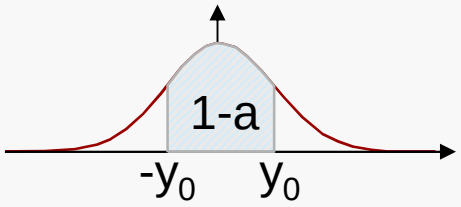
La variable centrée réduite

$$Y = \frac{nf - np}{\sqrt{npq}} = \frac{f - p}{\sqrt{\frac{p(1-p)}{n}}}$$

suit alors la loi normale centrée réduite N(0, 1).

Pour un niveau de confiance α choisi, les tables de la loi normale donnent la valeur y_o telle, que

$$P(-y_0 < Y < y_0) = 1-\alpha$$



Si Y $\in$ (-y_0, y_0), alors

$$P\left(-y_0 < \frac{f-p}{\sqrt{\frac{p(1-p)}{n}}} < y_0\right) = 1-\alpha$$

Donc

$$p \in \left(f - y_0\sqrt{\frac{p(1-p)}{n}},\ f + y_0\sqrt{\frac{p(1-p)}{n}}\right)$$

comme

$$\sigma = \sqrt{np(1-p)} = \sqrt{\frac{n}{n-1}}\sqrt{nf(1-f)}$$

alors

$$\frac{p(1-p)}{n} = \frac{f(1-f)}{n-1}$$

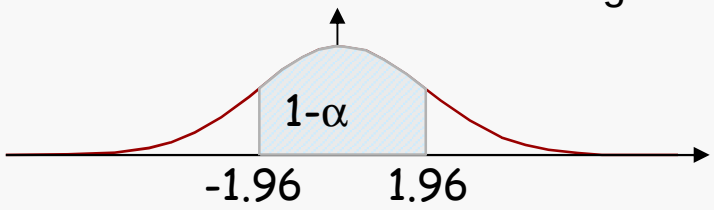
$$p \in \left(f - y_0\sqrt{\frac{f(1-f)}{n-1}},\ f + y_0\sqrt{\frac{f(1-f)}{n-1}}\right)$$

EXEMPLE 2: Un sondage effectue sur 100 personnes indique que 20 parmi ces personnes sont des grands fumeurs (ils fument plus que 20 cigarettes par jour). Au risque de 5%, quel est le pourcentage de grands fumeurs dans la population d'où est extrait l'échantillon?

Soit x – le nombre de grands fumeurs dans un échantillon de taille n.

f = x/n – fréquence d'apparition d'un grand fumeur dans l'échantillon

La variable x suit la distribution binomiale B(100, 0.2): $\mu = np = 20$ $\sigma = \sqrt{npq} = 4$.



n est grand → distribution binomiale peut être approximée par la loi normale N(20, 4).

Pour un niveau de confiance $\alpha = 5\ \%$
les tables de la loi normale donnent : $y_0 = 1.96$

$$P(-y_0 < Y < y_0) = 1 - \alpha$$

Donc, au risque 5 % , l'intervalle de confiance du pourcentage de grands fumeurs dans la population est :

$$p \in \left(f - y_0 \sqrt{\frac{f(1-f)}{n-1}}; f + y_0 \sqrt{\frac{f(1-f)}{n-1}} \right)$$

$$p \in \left(0.2 - 1.96 \sqrt{\frac{0.2 \cdot 0.8}{99}}; 0.2 + 1.96 \sqrt{\frac{0.2 \cdot 0.8}{99}} \right) \Rightarrow \boxed{p \in (0.192; 0.208)}$$

Intervalle de confiance d'une moyenne.

On étudie un paramètre **X** dans une population **P**.

Soit: $\bar{x}$ – la moyenne de statistique **X** dans un échantillon de taille **n,**
S^2 – sa variance.

QUESTION:

$\bar{x}$ est un estimateur ponctuel de μ.
Au niveau de confiance α, dans quelles limites est contenue la vraie valeur de μ ?

L'expérience montre qu'en pratique, quelle que soit la loi suivie par la variable **X**, les distributions d'échantillonnage suivent la loi normale.

Donc, pour la distribution des moyennes, la variable centrée réduite Y

$$Y = \frac{\bar{x} - \mu}{S / \sqrt{n-1}}$$

suit la loi normale centrée réduite N(0,1).

L'intervalle de confiance de la moyenne μ de la population est donc :

$$\mu \in \left(\bar{x} - y_0 \frac{S}{\sqrt{n-1}} ; \bar{x} + y_0 \frac{S}{\sqrt{n-1}} \right)$$

EXEMPLE 3: On a effectué 50 mesures du diamètre d'une sphère.
La moyenne des résultats a été $\overline{x}$ = 4.38 mm, écart type s = 0.06 mm.
Au niveau de confiance de 99 %, entre quelles limites est contenue la vraie valeur de ce diamètre?

Soit **x** – le résultat d'une mesure.
Comme le nombre de mesures effectuées est grand (n = 50), on peut supposer que la variable **x** est distribuée normalement.

Pour un niveau de confiance de **1 %** les tables de la loi normale donnent : $\mathbf{y_o = 2.58}$

Donc, au risque 1 % , la vraie valeur du diamètre de la sphère (moyenne d'un nombre infini de mesures) est contenue dans l'intervalle :

$$\mu \in \left(\bar{x} - y_0 \frac{S}{\sqrt{n-1}} ; \bar{x} + y_0 \frac{S}{\sqrt{n-1}} \right)$$

$$\mu \in \left(4.38 - 2.58 \frac{0.06}{\sqrt{50-1}} ; 4.38 + 2.58 \frac{0.06}{\sqrt{50-1}} \right)$$

$$\mu \in (4.36; 4.40)$$

EXEMPLE 4: Un échantillon de n = 17 trous effectuées avec une perforeuse montre le diamètre moyen $\bar{x}$ = 3.47 mm avec un écart type s = 0.05 mm. Entre quelles limites se situent les diamètres de 95% de trous faites avec cette machine?

Soit **x** – le résultat d'une mesure.
Comme le nombre de mesures effectuées est petit (n = 17), nous allons supposer que la variable **x** suit la loi de Student à **ν** = 17-1 =16 degrés de liberté.

Pour **ν** = 16 et le niveau de confiance de **5 %** les tables de la loi de Student donnent : $\mathbf{t_o = 2.12}$

Donc, au risque 5 % , la vraie valeur du diamètre des trous (moyenne d'un nombre infini de mesures) est contenue dans l'intervalle :

$$\mu \in \left(\bar{x} - t_0 \frac{S}{\sqrt{n-1}}; \bar{x} + t_0 \frac{S}{\sqrt{n-1}} \right)$$

$$\mu \in \left(3.47 - 2.12 \frac{0.05}{\sqrt{17-1}}; 3.47 + 2.12 \frac{0.05}{\sqrt{17-1}} \right)$$

$$\mu \in (3.44; 3.50)$$

REMARQUE:
utilisation de la loi normale (y_0 = 1.96) sous-estime la largeur de l'intervalle :

$$\mu \in (3.45; 3.49)$$

Intervalle de confiance de somme des moyennes.

Supposons que nous disposons de deux (ou plus) échantillons pour mener l'étude d'un paramètre **X**. Soit:

$\bar{x}_1$, S_1^2 – la moyenne et la variance de statistique **X** dans l'échantillon de taille $\mathbf{n_1}$,

$\bar{x}_2$, S_1^2 – la moyenne et la variance de statistique **X** dans l'échantillon de taille $\mathbf{n_2}$.

QUESTION:

$\bar{x}_1 \pm \bar{x}_2$ est un estimateur ponctuel de la somme (différence) entre les moyennes.
Au niveau de confiance α, dans quelles limites est contenue sa vraie valeur ?

Si seulement les échantillons sont grands, on peux les modéliser par la loi normale.

L'intervalle de confiance de la somme (différence) de moyennes est donc:

$$\left(\bar{x}_1 \pm \bar{x}_2 - y_0 \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}};\ \bar{x}_1 \pm \bar{x}_2 + y_0 \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} \right)$$

Si l'écart type σ_ι de populations dont les échantillons sont issus est connu , alors

$$\left(\bar{x}_1 \pm \bar{x}_2 - y_0 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ \bar{x}_1 \pm \bar{x}_2 + y_0 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

EXEMPLE 5: La force électromotrice des piles produites par une compagnie est normalement distribuée avec, en moyenne, E = 45.1 V et écart type **s** = 0.04 V.

Si on relie 4 piles en série, quelle est, au niveau de risque α = 5%, la gamme de tensions délivrée par cette batterie ?

Si les piles sont reliées en série, alors

$$E = E_1 + E_2 + E_3 + E_4 = 4 \cdot 45.1 = 180.4\,V$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_3^2}{n_3} + \frac{\sigma_4^2}{n_4}} = \sqrt{4 \cdot (0.04)^2} = 0.08\,V$$

Pour un niveau de confiance **α = 5 %** les tables de la loi normale donnent : **y_o = 1.96**.
Donc, l'intervalle de confiance de la tension délivrée par la batterie est:

$$E \in \left(E - y_0 \sqrt{\sum_{i=1}^{4} \frac{\sigma_i^2}{n_i}};\; E + y_0 \sqrt{\sum_{i=1}^{4} \frac{\sigma_i^2}{n_i}} \right)$$

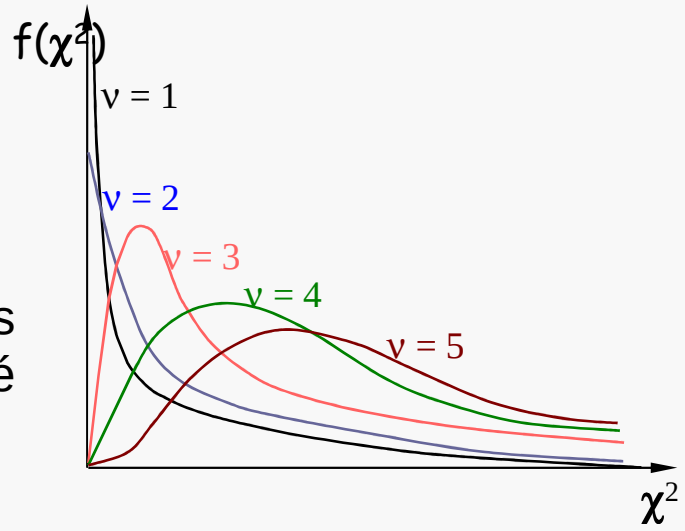
$$E \in (180.4 - 1.96 \cdot 0.08;\; 180.4 + 1.96 \cdot 0.08)$$

$$E \in (180.24;\; 180.56)$$

Intervalle de confiance d'une variance.

En étudiant les variances S^2 des paramètres, d'habitude on ne décrit pas directement la distribution de S^2 mais la distribution d'un paramètre relatif χ^2

$$\chi^2 = \frac{nS^2}{\sigma}$$

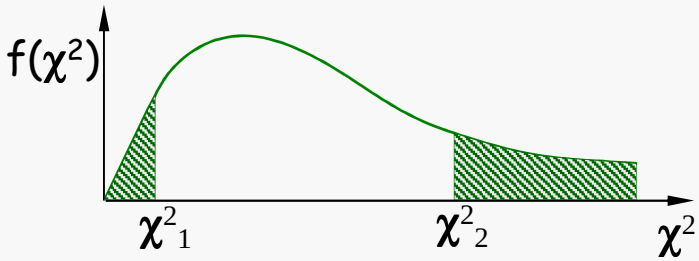


Ce paramètre suit la distribution asymétrique dont la forme dépend de la taille **n** des échantillons retirées de la population (du degré de liberté du système $\nu = n-1$)

Pour un risque α les tables de la loi χ^2 donnent les valeurs de paramètres χ^2_1 et χ^2_2 tels, que $P\left(\chi_1^2 < \chi < \chi_1^2\right) = 1-\alpha$

ce qui mène à

$$\frac{S\sqrt{n-1}}{\chi_2^2} < \sigma < \frac{S\sqrt{n-1}}{\chi_1^2}$$

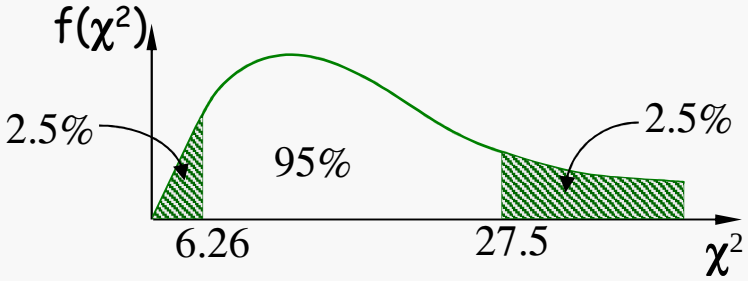


EXEMPLE 6: L'écart type de la durée de vie d'un échantillon de 15 ampoules choisies dans un lot de 1000 ampoules est de 2.40 h.
Entre quelles limites est contenu écart type de la durée de vie du lot entier? Donner la réponse au niveau de 5% de risque.

Pour **n** = 15, ν = **n** – 1 = 14

$\chi^2_{0.025}$ = 6.26

$\chi^2_{0.975}$ = 27.5



Donc, l'écart type du lot d'ampoules est contenu entre:

$$\frac{S\sqrt{n-1}}{\chi^2_2} < \sigma < \frac{S\sqrt{n-1}}{\chi^2_1}$$

$$\frac{2.4\sqrt{15}}{27.5} < \sigma < \frac{2.4\sqrt{15}}{6.26}$$

$$1.83 < \sigma < 3.84$$