Estimation

Population parameter (θ)

Estimator:

$$\hat{\theta} = \hat{\theta}_n = \hat{\theta}(y_1, y_2, \ldots, y_n)$$

- **Depends on the elements of sample** $(y_1, y_2, \ldots, y_n)$
- **Random/probability variable**

Sample characteristics

Population $(\mu, \sigma^2) \longrightarrow$ **IID sample (n elements)**

1. Expected value of the sample mean:

$$E(\bar{y}) = E\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n}E\left(\sum y_i\right) = \frac{1}{n}\left[E(y_1 + y_2 + \ldots + y_n)\right] =$$

$$= \frac{1}{n}\left[E(y_1) + E(y_2) + \ldots + E(y_n)\right] = \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot \mu = \mu$$

$$\boxed{E(X + Y) = E(X) + E(Y)}$$

$$\boxed{E(aX) = aE(X)}$$

$E(\bar{y}) = \mu \longrightarrow$ **Unbiased estimation**

Sample characteristics

2. Variance of the sample mean:

$$Var(\bar{y}) = Var\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n^2}\left[Var(y_1) + Var(y_2) + \ldots + Var(y_n)\right] =$$

$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \ldots + \sigma^2\right] = \frac{\cancel{n}\sigma^2}{n^{\cancel{2}}} = \frac{\sigma^2}{n} = \sigma_{\bar{y}}^2$$

$$\boxed{Var(aX) = a^2 Var(X)}$$

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)}$$

$$\boxed{Var(X+Y) = Var(X) + Var(Y)} \longrightarrow \text{IID sample}$$

Uncertainty multipliers

1. $1-\alpha = 90\%$

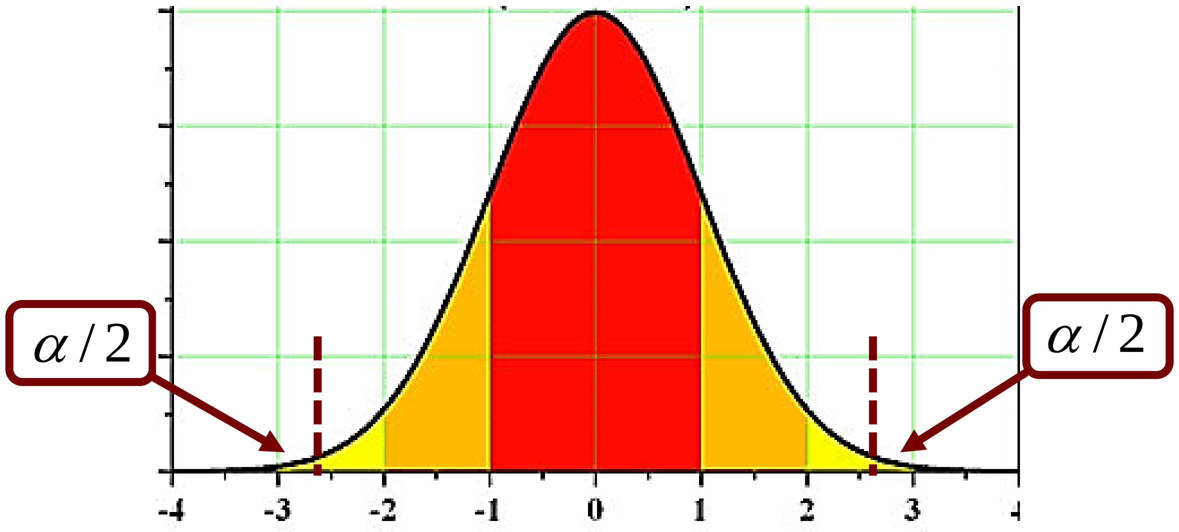
   $z_{0,95} = 1,645$

2. $1-\alpha = 95\%$

   $z_{0,975} = 1,960$

3. $1-\alpha = 99\%$

   $z_{0,995} = 2,576$



$\Phi(1) = 0,8413 \quad z = 1 \quad 1-\alpha = 68,27\%$

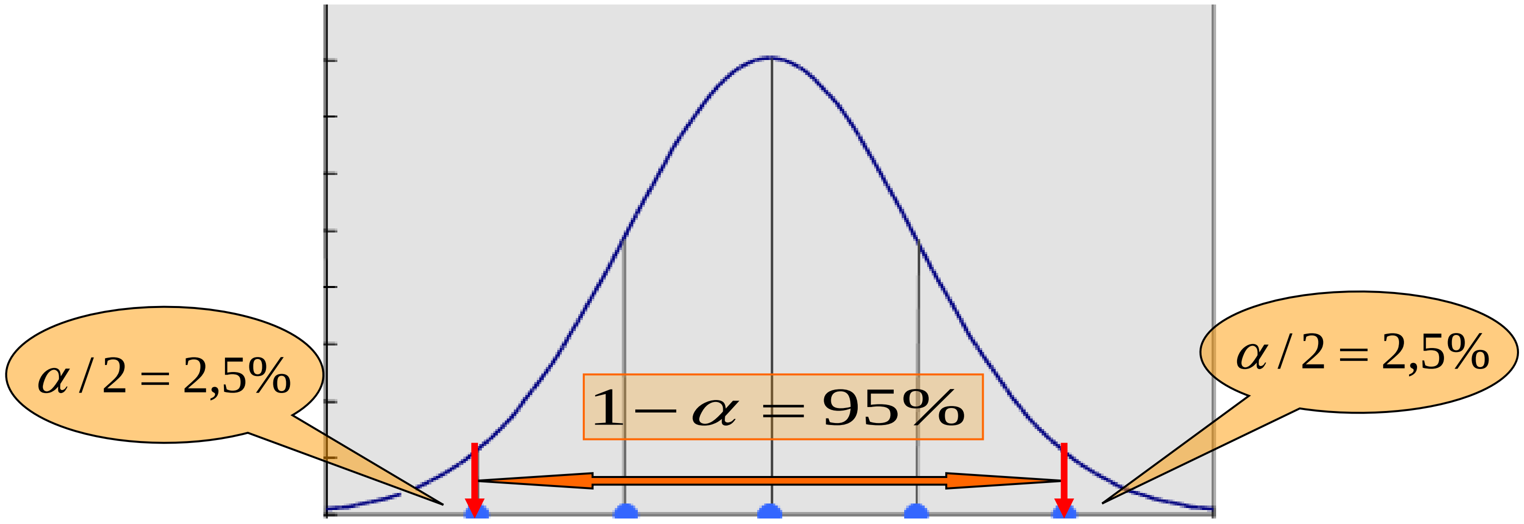
$\Phi(2) = 0,9772 \quad z = 2 \quad 1-\alpha = 95,45\%$

$\Phi(3) = 0,9987 \quad z = 3 \quad 1-\alpha = 99,73\%$

Estimation of expected value from IID sample

Normal distribution (Y), population sd (σ)

$$\bar{y} \sim N(\mu, \sigma/\sqrt{n})$$



$$\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \mu \qquad \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} = \mu + \boxed{z_{1-\alpha/2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Estimation from one sample:

$$\alpha = 5\% \qquad \bar{y} = 3{,}1$$

$$3{,}1 - 1{,}96 \cdot 0{,}06 < \mu < 3{,}1 + 1{,}96 \cdot 0{,}06$$

$$3{,}1 - 0{,}1176 < \mu < 3{,}1 + 0{,}1176$$

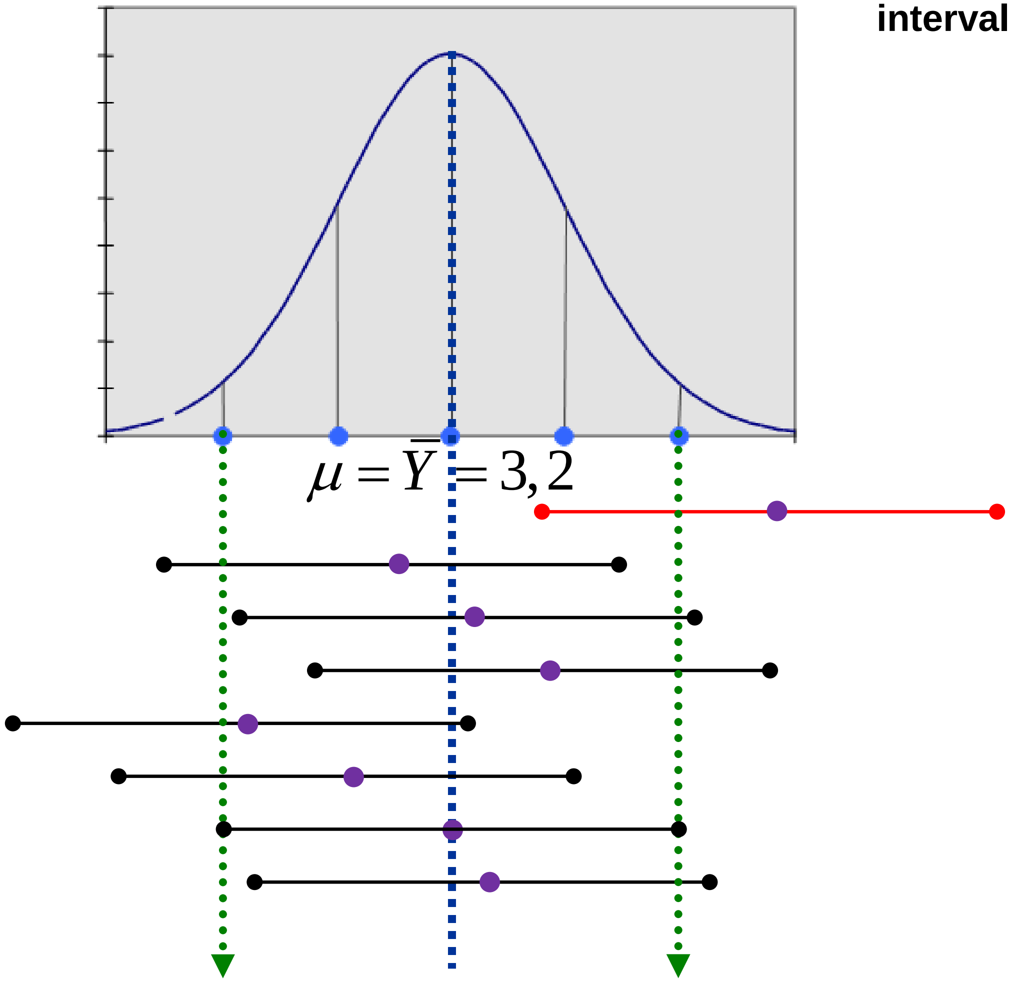
$$2{,}9824 < \mu < 3{,}2276$$

Resampling confidence intervals

$$3{,}0824 < \bar{y} < 3{,}3176$$

95% of all possible 100 elements IID sample means fall within this interval.

$$\Delta_{\bar{y}} = 0{,}1176$$



Sample characteristics

Population (μ, σ^2) $\longrightarrow$ **IID sample (n=100)**

$$s^{*2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

s^{*2} ***expected value:***

$$E\left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] =$$

$$\frac{1}{n} E\left[\sum \left(y_i^2 + \bar{y}^2 - 2 \cdot y_i \cdot \bar{y}\right)\right] = \frac{1}{n} E\left[\sum y_i^2 + \sum \bar{y}^2 - 2 \cdot \sum y_i \cdot \bar{y}\right]$$

Sample characteristics

$$\frac{1}{n}E\left[\sum y_i^2+\sum \bar{y}^2-2\cdot\overbrace{\sum y_i}^{n\cdot\bar{y}}\cdot\bar{y}\right]=\frac{1}{n}E\left[\sum y_i^2+\sum \bar{y}^2-2\cdot n\cdot\bar{y}^2\right]=$$

$$\frac{1}{n}\left[E\left(\sum y_i^2\right)+E\left(\sum \bar{y}^2\right)-2\cdot n\cdot E\left(\bar{y}^2\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\left[E\left(y_i^2\right)+E\left(\bar{y}^2\right)-2\cdot E\left(\bar{y}^2\right)\right]=$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[E\left(y_i^2\right)-E\left(\bar{y}^2\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\left[\left(\sigma^2+\mu^2\right)-\left(\frac{\sigma^2}{n}+\mu^2\right)\right]$$

$$\boxed{E\left(Y^2\right)=Var(Y)+\mu^2}$$

Sample characteristics

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left(\sigma^2+\mu^2\right)-\left(\frac{\sigma^2}{n}+\mu^2\right)\right] = \frac{1}{n}\left(n\cdot\sigma^2+\cancel{n\cdot\mu^2}-\cancel{n}\cdot\frac{\sigma^2}{\cancel{n}}-\cancel{n\cdot\mu^2}\right)$$

$$=\frac{1}{n}\left(n\cdot\sigma^2-\sigma^2\right) \qquad =\frac{\sigma^2}{n}(n-1) \qquad =\sigma^2-\frac{\sigma^2}{n} \qquad \neq\sigma^2$$

$$E\left(s^2\right)=E\left[\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}{n-1}\right]=\sigma^2$$

Sample characteristics

The (not corrected) empirical variance from the (IID) sample is a biased estimate of the population variance.

$$E\left(s^{*2}\right) = E\left[\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n}\right] = \sigma^2 - \boxed{\frac{\sigma^2}{n}} \neq \sigma^2$$

Bias

$$Bs\left(s^{*2}\right)$$

The (corrected) empirical variance from the (IID) sample is an unbiased estimate of the population variance.

$$E\left(s^2\right) = E\left[\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}\right] = \sigma^2$$

Estimation of expected value from IID sample

1. **Normal distribution (Y), population sd (*small sample*)**

$$\boxed{Int_{1-\alpha}(\mu) = \bar{y} \pm z_{1-\alpha/2} \cdot \sigma_{\bar{y}}}$$

$$\boxed{\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}}$$

2. **Normal distribution (Y), sample sd (*small sample*)**

$$\boxed{Int_{1-\alpha}(\mu) = \bar{y} \pm t_{1-\alpha/2} \cdot s_{\bar{y}}}$$

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1}$$

$$\frac{\bar{y}-\mu}{s_{\bar{y}}} = t \quad ahol \quad s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

3. **Asymptotic case**

$$\bar{y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad \textbf{or} \qquad \bar{y} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

Sample size

$$Int_{1-\alpha}(\mu) = \bar{y} \pm \boxed{z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}} = \bar{y} \pm \boxed{\Delta_{\bar{y}}}$$

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{\Delta_{\bar{y}}} \right)^2$$

$z_{1-\alpha/2}$	**Uncertainty multiplier**
σ	**Standard deviation**
$\Delta_{\bar{y}}$	**Margin of error**

Estimation properties

Unbiased

$$E(\hat{\theta}) = \theta$$

Measure of bias

$$Bs(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Efficiency

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

MSE

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bs^2(\hat{\theta})$$

Consistency

$$\lim_{(n \to \infty)} E\left(\hat{\theta}_n\right) = \theta \quad \lim_{(n \to \infty)} Var\left(\hat{\theta}_n\right) = 0$$

Proportion estimation (IID)

- **Population poportion:** $$P = \frac{K}{N}$$

K: the number of elements in the population with a given property

- **Estimator:**

$$\hat{P} = p = \frac{k}{n}$$

Proportion estimation (IID)

- **Characteristics:**

$$E(p) = P$$ **(unbiased)**

$$Var(p) = \sigma_p^2 = \frac{P(1-P)}{n}$$

Estimation:

Unbiased estimator

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

Biased estimator

$$s_p^2 = \frac{p(1-p)}{n}$$

For large samples it is acceptable

Proportion estimation (IID)

- **Distribution:** binomial

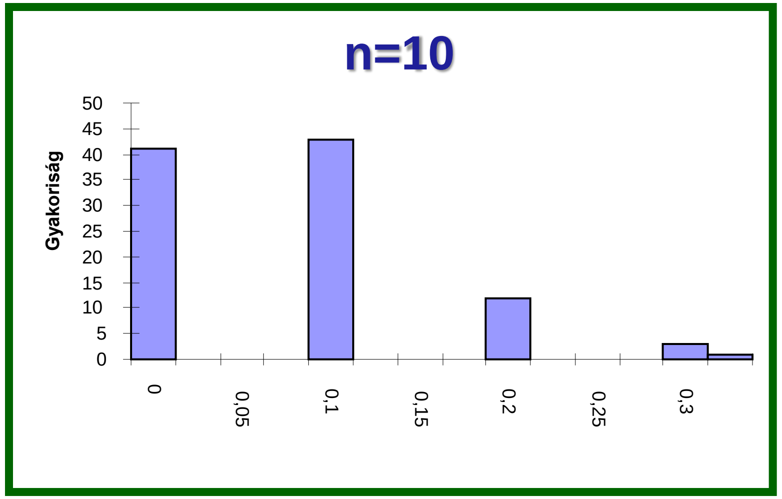
- **Can be approximated by a normal distribution for a large enough sample:**

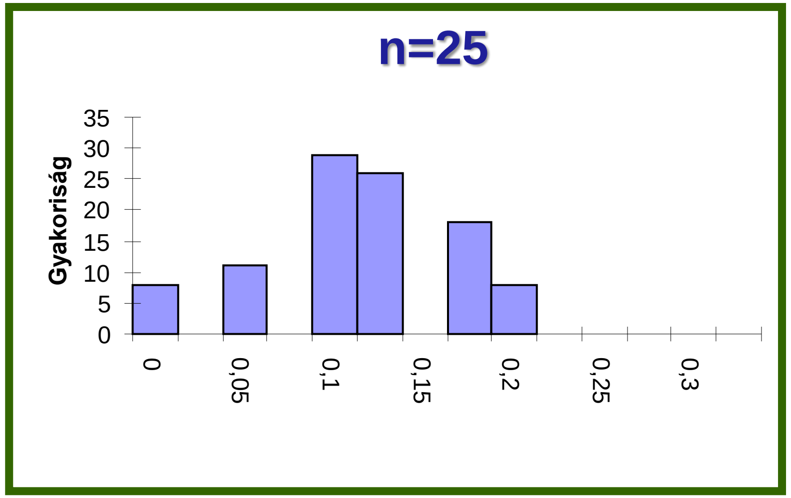
$$z = \frac{p - P}{s_p} \approx N(0,1)$$

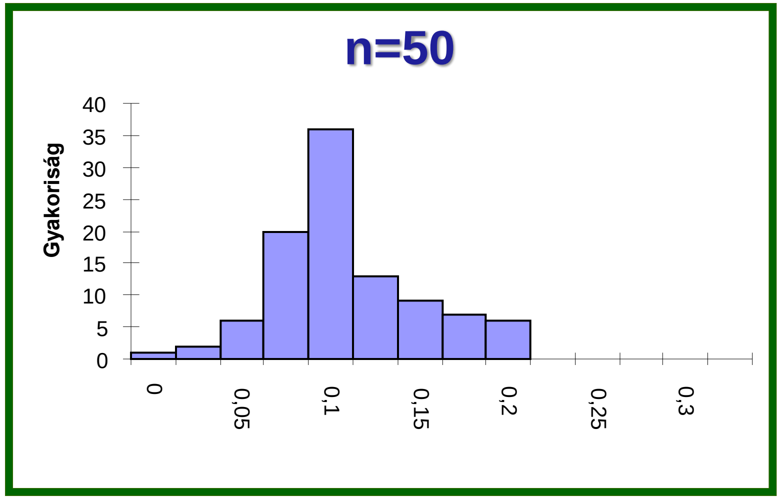
- **Neccessary number of observations:**

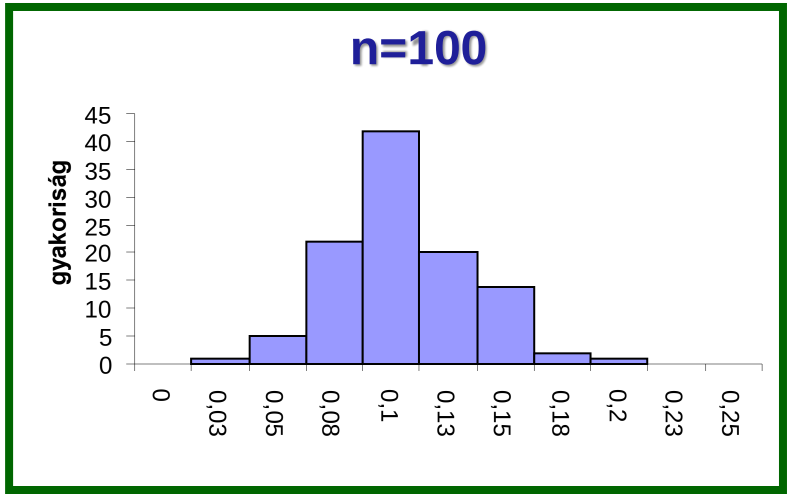
$$\min\{nP, n(1-P)\} \geq 10$$

Distribution of proportion estimation $(P = 0{,}1)$









Proportion estimation (IID)

- **Confidence interval**

$$\overset{\text{point}}{p} \pm \underbrace{z_{1-\alpha/2}} \cdot \underbrace{\sqrt{\frac{p\,(1-p)}{n}}}_{se} = p \pm z_{1-\alpha/2} \cdot s_p = p \pm \Delta_p$$

Sample size

$$z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} = \Delta_p$$

❖ **Sample size:**

max. 0,25

$$n = \frac{z^2_{1-\alpha/2} \cdot \boxed{P \cdot (1-P)}}{\Delta^2}$$

Variance (σ^2) estimation (IID)

Estimators

1)

$$s^{*2} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n} \qquad E(s^{*2}) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$$

(biased estimator)

2)

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} \qquad E(s^2) = \sigma^2$$

(unbiased estimator)

Variance (σ^2) estimation (IID)

- **Point estimation:**

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- **Interval estimation:**

Condition: **distribution of the population is normal**

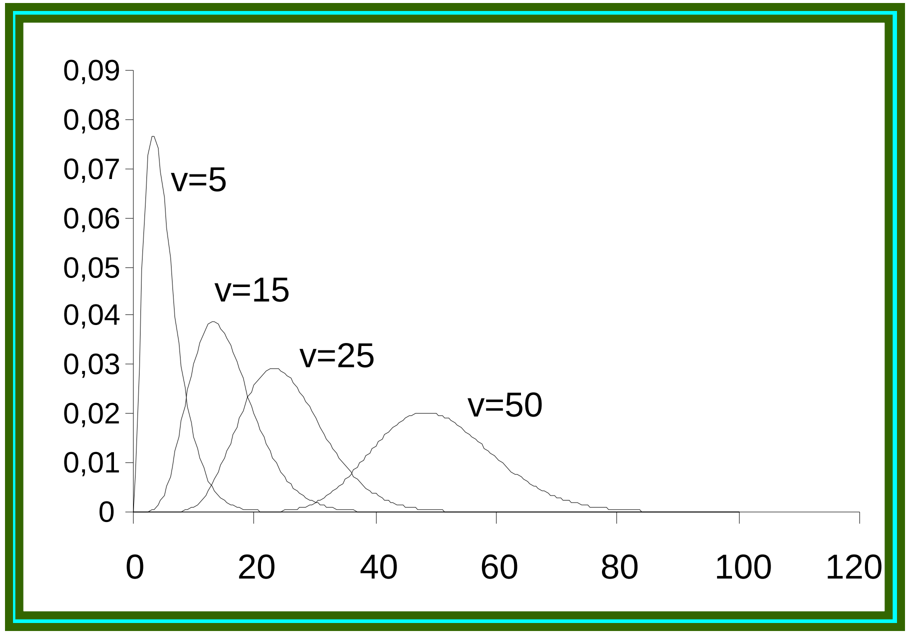
The $\frac{(n-1)s^2}{\sigma^2}$ **variable is probability variable which follows** χ^2 **distribution with** $\nu = n-1$ **degrees of freedom**

Chi-square distribution

n independent probability variables with standard normal distribution:

$$Z_1, Z_2, \ldots\ldots, Z_n$$

$$U = Z_1^2 + Z_2^2 + \ldots\ldots + Z_n^2 = \sum Z_i^2$$ **Distribution of U probability variable:** χ_n^2



Characteristics:

- **Right skewed**
- **Approximates the normal distribution by increasing the sample size.**
- **Values between 0 and + infinite.**
- **Expected value** $E(\chi^2) = n$
- **Variance** $Var(\chi^2) = 2n$

Variance estimation

- **Probability:**

$$P\left(\chi^2_{\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

- **Confidence interval for:** σ^2

$$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(\nu)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}(\nu)}\right) = 1-\alpha \qquad \nu = n-1$$

Not symmetric to the point estimate!

Estimation of amount of value

- **Population:** $Y' = \sum_{i=1}^{N} Y_i = N \cdot \bar{Y}$

- **Estimatios:** $\hat{Y}' = N \cdot \bar{y}$

$$E\left(\hat{Y}'\right) = E\left(N \cdot \bar{y}\right) = N \cdot E\left(\bar{y}\right) = N \cdot \bar{Y} = Y'$$

- **Confidence interval:** $N \cdot \left(\bar{y} \pm \Delta_{\bar{y}}\right)$

Estimation of K in population

- **Sample proportion:** p
- **Point estimation:** $N \cdot p$
- **Confidence interval:** $N \cdot \left(p \pm \Delta_p \right)$

Simple random sample

- **Condition:** ***list about the population***

- **Characteristics:**
  - **N is important**
  - **The elements are not independent**
  - **Why? Because of no replacement**

- **Sample size should be large**

  In this way we use (standard) normal distribution.

Variance estimation

$$\boxed{Var\ (\bar{y})} = Var\left(\frac{1}{n}\sum y_i\right) =$$

$$= \frac{1}{n^2}\left[Var(y_1) + Var(y_2) + \ldots + Var(y_n) + \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(y_i, y_j)\right] =$$

$$i \neq j$$

$$= \frac{1}{n^2}\left[n \cdot \sigma^2 + n \cdot (n-1)\left(-\frac{\sigma^2}{N-1}\right)\right] = \frac{\sigma^2}{n}\left[1 + (n-1)\left(-\frac{1}{N-1}\right)\right] =$$

$$\boxed{= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

$$\boxed{Var(aX) = a^2 Var(X)} \quad \boxed{Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)}$$

Variance estimation

$$Var(\bar{y}_{EV}) = \sigma^2_{\bar{y}(EV)} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \cong \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) \cong \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$$

- Compared to iid it is more accurate
- $\left(\frac{n}{N}\right)$ proportion matters, except for large population

Standard error

$$\sigma_{\bar{y}(EV)} \cong \frac{\sigma}{\sqrt{n}} \underbrace{\sqrt{1-\frac{n}{N}}}_{\text{Correction factor}}$$

Correction factor does not matter if the n/N is less than 1%

Confidence interval

$$\bar{y} \pm \Delta_{\bar{y}}$$

$$\Delta_{\bar{y}} = z_{1-\alpha/2} \cdot \sigma_{\bar{y}(EV)} = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

For sample sd:

$$\Delta_{\bar{y}} = z_{1-\alpha/2} \cdot s_{\bar{y}(EV)} = z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Number of observations

$$\Delta = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n_{EV}}} \sqrt{1 - \frac{n_{EV}}{N}}$$

Thus:

$$n_{EV} = \frac{z_{1-\alpha/2}^{2} \cdot \sigma^{2}}{\frac{z_{1-\alpha/2}^{2} \cdot \sigma^{2}}{N} + \Delta^{2}}$$

IID and simple random sample

- **Expected value:**

$$E(\bar{y}_{FAE}) = \mu \qquad\qquad E(\bar{y}_{EV}) = \mu$$

- **Variance:**

$$Var(\bar{y}_{FAE}) = \frac{\sigma^2}{n} \quad \geq \quad Var(\bar{y}_{EV}) \cong \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

$$\frac{Var\left(\bar{y}_{FAE}\right)}{Var\left(\bar{y}_{EV}\right)} \approx \frac{1}{1 - \dfrac{n}{N}} \geq 1$$

Estimation of P in simple random sample

$$p \pm z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n}{N}} = p \pm \Delta_p$$