If n is large and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by normal distribution with std normal r.v.:

$$Z = \frac{X - np}{\sqrt{npq}}$$

For X binomial r.v.: $np > 5$, $nq > 5$

$$E[X] = m_x = np$$

$$\sigma_x = \sqrt{npq}$$

Ex) A fair coin is tossed 500 times. Find the probability that the number of heads will not differ from 250 by more than 10.

$$P(240 < \underbrace{X_b}_{\text{binomial}} < 260) \approx P(239.5 < \underbrace{X_n}_{\substack{\text{normal with} \\ \text{continuity correction}}} < 260.5)$$

$$= P\left(\frac{239.5 - 250}{11.18} < Z_n < \frac{260.5 - 250}{11.8}\right)$$

where $m = np = (500)\left(\frac{1}{2}\right) = 250$ and

$$\sigma = \sqrt{npq} = \sqrt{500\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} \simeq 11.18$$

$$= P(-0.94 < Z < 0.94) = 2P(0 < Z < 0.94)$$

$$= 2A(0.94) = 2(0.3264) = 0.6528$$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent random variables from some probability distribution with mean μ and variance σ^2. Then if

$S_n = X_1 + X_2 + \ldots + X_n$ then: $E[S_n] = E[X_1] + \ldots + E[X_n] = \mu + \ldots + \mu = n\mu$

$$\lim_{n \to \infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du$$

For sample means:

if n large ($n > 30$) then distribution of sample means $\{\bar{x}\}$ for <u>all</u> possible samples of size n is approximately normal with mean $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

what this says: $E[\bar{x}] = \mu$ and $\sigma_{\bar{x}}$ decreases with sample size. So values of sample mean cluster more and more closely around population mean as n increases.

confidence interval the probability $(1-\alpha)$ that the interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. $\alpha \in (0,1)$

An interval estimate of θ is an interval of the form $\hat{\theta}_1 < \theta < \hat{\theta}_2$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are appropriate values of r.v. θ s.t:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha \quad \text{for } \alpha \in (0,1).$$

Note: like point estimates, interval estimates of a given parameter are not unique.

The $z_{\alpha/2}$ value

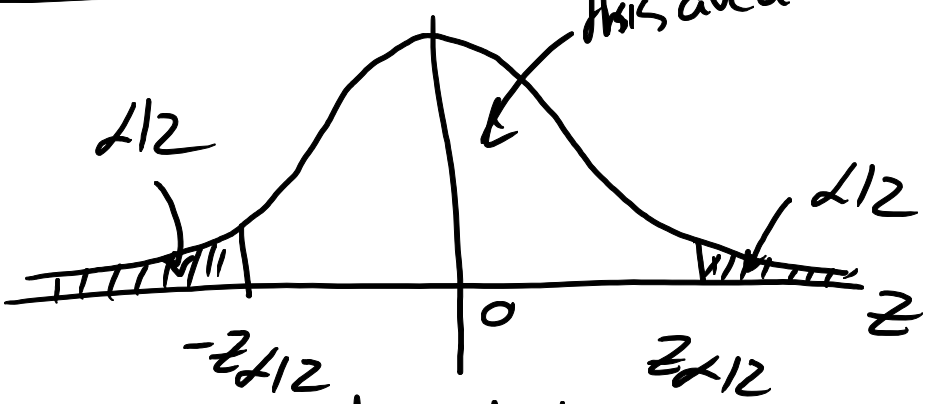
place area $\alpha/2$ in each tail.

If we place area $\alpha/2$ in each tail and if $z_{\alpha/2}$ is the value of z such that area $\alpha/2$ will lie to its right, then the confidence

interval with confidence coefficient $(1-\alpha)$ (the probability that an interval estimator encloses a population parameter) is

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$

$\rightarrow \bar{x} - z_{\alpha/2}\sigma_{\bar{x}} < m < \bar{x} + z_{\alpha/2}\sigma_{\bar{x}}$



α-dependent # which gives area $1-\alpha$.

$$P(|z| < z_{\alpha/2}) = 1-\alpha = P(-z_{\alpha/2} < z < z_{\alpha/2})$$

where $z = \dfrac{\bar{x}-m}{\sigma/\sqrt{n}} = \dfrac{\bar{x}-m_{\bar{x}}}{\sigma_{\bar{x}}}$

$$|z| = \frac{|\bar{x}-m|}{\sigma/\sqrt{n}} = \frac{|\bar{x}-m|}{\sigma_{\bar{x}}}$$

$$\Rightarrow P\left(|\bar{x}-m| < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\Rightarrow$ <u>Thm</u> If $\bar{x}$, the mean of a random sample of size n from a <u>normal</u> population with known variance σ^2 is to be used as an <u>estimator</u> of the mean of the population,

the probability is $(1-\alpha)$ that the error will be less than $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. If the population not normal, $n \geq 30$ needed for CLT.

Ex) A team of factory contractors uses the mean of a random sample of size $n=150$ to estimate the population mean of a factory product. If based on experience $\sigma = 6.2$ is known (population std. deviation), what can be asserted with 0.99 probability about the max error of their estimate? Want to bound $|\bar{x}-u|$ with 99% probability.

$\Rightarrow n=150,\ \sigma=6.2,\ \alpha=0.01 \Rightarrow \frac{\alpha}{2}=0.005$

$$P(|z| < z_{\alpha/2}) = 1-\alpha = 0.99 \quad \Big| \quad z=\frac{\bar{x}-u}{\sigma/\sqrt{n}}$$

$1-0.01=0.99$

$$\Rightarrow P(-z_{\alpha/2} < z < z_{\alpha/2}) = 0.99$$

$$\Rightarrow 2P(0 < z < z_{\alpha/2}) = 0.99$$

$$\Rightarrow P(0 < z < z_{\alpha/2}) = \frac{0.99}{2} = 0.495 \simeq 0.4949$$

$\Rightarrow \underbrace{z_{\alpha/2}}_{z_{0.005}} = 2.57$ from table

Thus, we plug into: $P(|Z| < z_{\alpha/2}) = 1-\alpha$

$$P\left(|\bar{x}-\mu| < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha = 0.99 \quad \Rightarrow \quad P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = 1-\alpha$$

this is the error term

$$\Rightarrow z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} = 2.57\,\frac{6.2}{\sqrt{150}} \simeq 1.30$$

Thus the team can assert with probability 0.99 that the estimation error is less than 1.30.

Note: $P\left(|\bar{x}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ can be written as: (confidence interval for μ).

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Thm If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, then: (or $n \geq 30$, from arbitrary population)

$$\left\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for the mean of the population.

Ex) If a random sample of size $n=20$ from a normal population with variance $\sigma^2=225$ has the mean $\bar{x}=64.3$, construct a 95% confidence interval for the population mean μ.

$= P(-z_{\alpha/2} < z < z_{\alpha/2})$

$\Rightarrow n=20,\ \bar{x}=64.3,\ \sigma=15$; $P(|z| < z_{\alpha/2}) = 1-\alpha$

95% interval $\Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$2P(0 < z < z_{\alpha/2}) = 0.95 \Rightarrow P(0 < z < z_{\alpha/2}) = 0.475$

$\Rightarrow z_{\alpha/2} = 1.96$ (from table)

Thus, we get the interval: ✓ 95% confidence interval for μ

$$64.3 - 1.96\frac{15}{\sqrt{20}} < \mu < 64.3 + 1.96\frac{15}{\sqrt{20}}$$

$$\Rightarrow \{57.7 < \mu < 70.9\} \quad \left| z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \right.$$

Ex) Use the following data (random sample) to construct an approximate 95% confidence interval for the mean of the population sampled (assuming its normal since n small).

$n=7$

$$\{10, 12, 9, 6, 4, 3, 2\}$$

$$\bar{x} = \frac{\Sigma x}{n} = 6.57 \ ; \quad s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \approx 3.9$$

$z_{.025} = 1.96$ (from table, as before)

use s as estimate for σ.

$\Rightarrow 6.57 - 1.96\frac{3.9}{\sqrt{7}} < m < 6.57 + 1.96\frac{3.9}{\sqrt{7}}$

$\Rightarrow$ Very rough estimate; better use t-distribution!

notice that this interval would be quite large since number of samples (n) is small. Using $\sigma \approx s$ not accurate for such small sample.

Estimation of differences between means

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

← for indep random samples from a normal population

has the standard normal distribution

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

when populations normal or $n, n_2 \geq 30$.

Sample Sum statistics

$$Y = X_1 + X_2 + \ldots + X_n$$

X_i, indep. identically distributed from population with mean m and variance σ^2

$$E[Y] = E[X_1] + \ldots + E[X_n] = m + \ldots + m = nm$$

$$Var[Y] = Var[X_1] + \ldots + Var[X_n] = \sigma^2 + \ldots + \sigma^2 = n\sigma^2$$

$$\Rightarrow std[Y] = \sigma\sqrt{n}$$

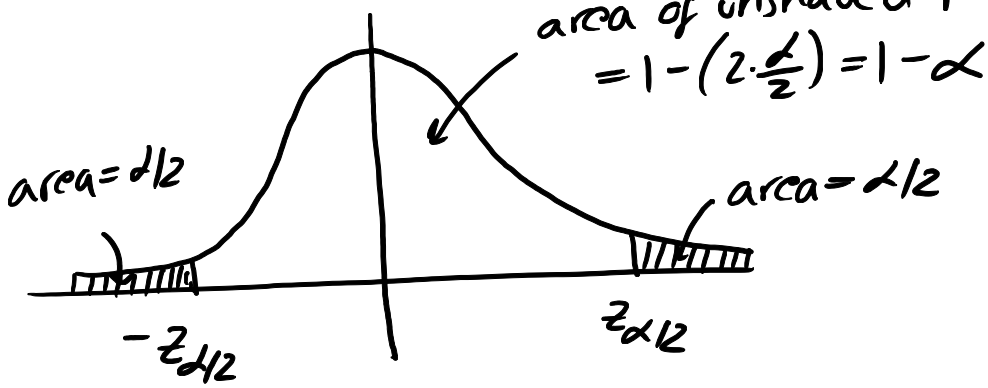
(I) Confidence intervals for unknown mean from general large populations

want: lower value $<$ population parameter (mean) $<$ upper value

(I-A) large scale estimation $n \geq 30$, σ known

When $n \geq 30$, we can make use of the CLT. In particular, no matter what the underlying population distribution is, the sample means $\bar{x}$ are approx. normal with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

We pick $z_{\alpha/2}$ s.t.



That is for a std normal r.v. z, we have:

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow P(|z| < z_{\alpha/2}) = 1-\alpha$$

Since in the case we are considering

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is std. normal (mean=0, std dev=1)

we have:

$$P\left(\underbrace{|\bar{x} - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{the error in estimation of } \mu}\right) = 1 - \alpha \quad ; \quad \alpha \in (0,1)$$

That is, the probability is equal to $(1-\alpha)$ that in the estimation of unknown μ by $\bar{x}$, the error is less than $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is a $(1-\alpha) \cdot 100\%$ confidence interval for μ.

(I-B) smaller scale estimation from a <u>normal population</u>, σ known

If $n < 30$ but the population is known to be normal with σ known, the same result applies.

Ex) If a random sample of size $n=20$

from a normal population with variance $\sigma^2 = 225$ has mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

$$n = 20, \quad \sigma = 15, \quad \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$$

$z_{\alpha/2} = z_{.025} = 1.96$, which we get from definition:

$1 - \alpha = 1 - 0.05 = 0.95$

$$P(|Z| < z_{0.025}) = .95 = 2P(0 < Z < z_{0.025})$$

$$\Rightarrow P(0 < Z < z_{0.025}) = \frac{.95}{2} = 0.475$$

$$\Rightarrow z_{0.025} = 1.96 \text{ (from table)}$$

(I-C) large sample estmation, σ unknown

when $n \geq 30$, CLT applies so the distribution of sample means is approximately normal.

In many cases, σ is not known (population std. dev.)

But if sample is large enough ($n \geq 30$), we can use s for σ where $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ sample std dev.

confidence interval for μ is: $\bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$

Ex) Suppose a large airline wants to estimate its average # of unoccupied seats per flight over the past year. To do this, 225 flights are selected randomly $\Rightarrow \bar{x} = 11.6$ seats and $s = 4.1$ seats

(Here σ unknown, sample std dev. was computed)

Estimate μ using a 90% confidence interval:

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

s used for σ. ok since n is large.

notice: we don't need to assume population is normal and we use s as estimate for σ since n is large.

$\alpha = 0.1 \Rightarrow \alpha/2 = 0.05$

$P(|z| < z_{\alpha/2}) = 1 - \alpha = 0.90$ $\quad \left| \; z_{\alpha/2} = z_{0.05} \right.$

$\Rightarrow P(0 < z < z_{\alpha/2}) = \frac{0.90}{2} = 0.45$

$z_{\alpha/2} = 1.64$ from table (can take $z_{\alpha/2} = 1.645$)

$\Rightarrow 11.6 - 1.64\frac{4.1}{\sqrt{225}} < \mu < 11.6 + 1.64\frac{4.1}{\sqrt{225}}$

is the confidence interval (of 90%)

(I-D) smaller scale estimation, σ generally unknown

($n < 30$, arbitrary population distribution)

use student-t distribution

The sample mean $\bar{x}$ follows the t distribution with mean μ and std. deviation $\frac{s}{\sqrt{n}}$.

(claimed to be discovered by a man in the Guinness brewery in Ireland, who published the result in 1908 under the pen name student).

use t distribution with mean $\bar{x}$, std. dev. $\frac{s}{\sqrt{n}}$.

t-distribution statistic is more variable than z since t contains two random quantities ($\bar{x}$ and s). The amount of variability depends on sample size n. For large n ($n \geq 30$), student t distribution approaches the normal distribution.

Ex) Suppose a drug company is testing a new fever reducing drug. The SIX ($n=6$) test patients sampled have a blood pressure increase of:

$\{1.7, 3.0, 0.8, 3.4, 2.7, 2.1\}$ points

when using the drug. Use this information to construct a 95% confidence interval for μ, the mean increase in blood pressure associated with the new drug for all patients in the population.

$\Rightarrow$ since n is small and we do not know σ (population std deviation) we use t-distribution.

$$\bar{x} = \frac{\sum x}{n} = \frac{13.7}{6} = 2.28$$

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{\sum (x-2.28)^2}{5} \approx .9017$$

The quantity $(n-1)$ is referred to as the # of degrees of freedom (df). Use

$$df = n-1 = 5$$

$$\Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$$

contrast with $z_{\alpha/2}$

$$\rightarrow \left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for the mean of the population.

$t_{0.025,5} \overset{n-1=df}{=} 2.571$ (from t-dist. table)

$$\Rightarrow \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 2.28 \pm (2.571)\left(\frac{.95}{\sqrt{6}}\right)$$

$$\approx 2.28 \pm 1.00$$

That is, we can be 95% confident that the mean increase in blood pressure associated with taking this new drug is between 1.28 and 3.28 points.

Note: for distinctly non-normal distributions t-statistic may not be accurate for constructing confidence intervals (use larger n or use a so called non-parametric method).

(II) large sample estimation of proportions

Suppose X is binomial r.v.

for large n, $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$ is approx. std normal

since $\mu_X = np$ and $\sigma_X = \sqrt{npq} = \sqrt{np(1-p)}$

Recall: $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1-\alpha$

substitute in the above value of z:

$$P\left(-z_{\alpha/2} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right) = 1-\alpha$$

Thm If X is a binomial random variable with parameters n and p, n large and $\hat{p} = \frac{X}{n}$ is a sample proportion then

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is an approximate $(1-\alpha)\cdot 100\%$ confidence interval for p.

Ex) In a random sample, 136 of 400 people given a flu vaccine experienced discomfort. Construct a 95% confidence interval for true proportion of people who will experience discomfort from the vaccine.

$\Rightarrow \quad n = 400, \quad \hat{p} = \frac{136}{400} = 0.34, \quad z_{.025} = 1.96$ from table

$\Rightarrow 0.34 - 1.96\sqrt{\frac{0.34(0.66)}{400}} < p < 0.34 + 1.96\sqrt{\frac{0.34(0.66)}{400}}$