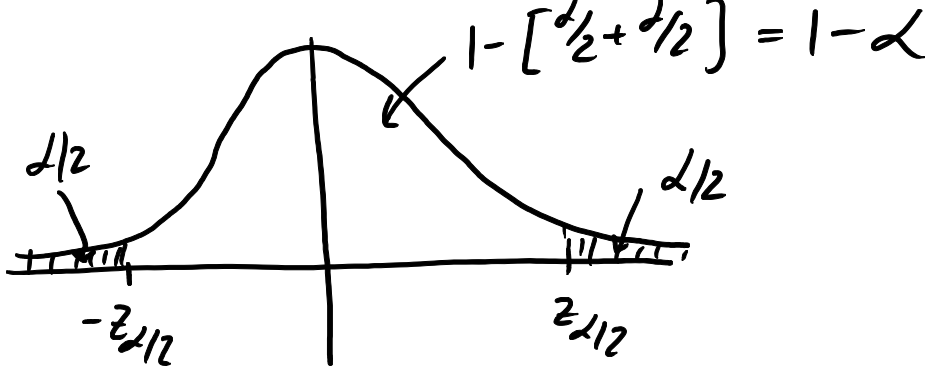
One mean estimation (large sample, small sample)
Two mean estimation (large sample, small sample)
Estimation of population proportion.

Review confidence intervals

want estimate on population parameter or proportion (e.g. μ, p).

use sample of size n to compute estimate.

Supply α-value. Then construct an <u>interval</u> which contains population parameter with probability $1-\alpha$.



$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1-\alpha = P(|z| < z_{\alpha/2})$$
$$= 2P(0 < z < z_{\alpha/2})$$

Where z is standard normal random variable.

When can we make use of this? large scale estimation

(1) large samples (large n). $n \geq 30$.

Then $\bar{x}$ is approx normal by CLT and

$\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ so that

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is std. normal

If σ (population std deviation) is not known, replace with $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ sample std deviation

(2) When population is known to be normal (then any samples drawn from it are normally distributed).

$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

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

Expand inequality $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \Rightarrow \bar{x} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$\Rightarrow \mu > \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

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

confidence interval for μ:

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

μ is contained in this interval with probability $1-\alpha$ for $\alpha \in (0,1)$. The smaller the value of α, the larger the interval.

Small scale estimation (small random sample)

$n < 30$, CLT does not apply

Assume we do not know that population is definitely normal. no basis to assume $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Replace std normal with student-t distribution

$z_{\alpha/2} \rightarrow t_{\alpha/2, n-1}$ $\quad \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim$ T-distribution

two parameters (α, degrees of freedom)

$df = n-1$ (which appears in $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$).

small sample confidence interval for μ:

$$\bar{X} - t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$

Ex) printer manufacturer tests $n=15$ printheads and calculates the following statistics:

$\bar{x} = 1.23$ mil chars $s = .27$ mil chars

before printhead fails

Form a 99% confidence interval for the mean number of characters printed.

$\Rightarrow n = 15 < 30$, know nothing about population distribution

$\Rightarrow$ use student-t distribution

$\alpha = 0.01 \Rightarrow \alpha/2 = 0.005$

$t_{0.005,14} = t_{\alpha/2,n-1} = 2.977$ (see table)

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) = 1.23 \pm 2.977\left(\frac{.27}{\sqrt{15}}\right)$$

$$= 1.23 \pm .21$$

$\Rightarrow (1.02 < \mu < 1.44)$

Thus, the manufacturer can be 99% sure that the mean print life is at least 1.02 min characters.

Note: as $n \to 30$, std-t distribution approaches std. normal.

Estimation of proportions

For X binomial r.v. and n large

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \to N(0,1) \quad \text{(std. normal)}$$

as $n \to \infty$. (Recall $\mu_X = np$; $\sigma_X = \sqrt{npq}$)

p success probability. $q = 1-p$ failure probability.

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

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

$$\Rightarrow \frac{X - np}{\sqrt{np(1-p)}} > -z_{\alpha/2} \Rightarrow X - np > -z_{\alpha/2}\sqrt{npq}$$

$$X > np - z_{\alpha/2}\sqrt{npq}$$

$$\Rightarrow \frac{X}{n} > p - z_{\alpha/2}\sqrt{\frac{npq}{n^2}} \Rightarrow \frac{X}{n} > p - z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

notice that $\frac{X}{n}$ is the proportion of successes in n samples.

Similarly, $X < np + z_{\alpha/2}\sqrt{npq}$ gives

$$\frac{X}{n} < p + z_{\alpha/2}\sqrt{\frac{pq}{n}} \Leftarrow \text{replace } \sqrt{\frac{pq}{n}} \text{ by } \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \downarrow \text{ by Law of Large Numbers}$$

$\hat{p}$

$\Rightarrow$ denote $\frac{x}{n}$ by $\hat{p}$ and write confidence interval for p:

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

$(1-\alpha)\cdot 100\%$ confidence interval for p. (p: population proportion of successes)

Ex) 136 of 400 people given a flu vaccine experienced discomfort. Construct 95% interval for true proportion of patients experiencing discomfort from the vaccine.

$n = 400$ (large scale, confidence interval applies)

$$\hat{p} = \frac{136}{400} = 0.34$$

95% CI $\Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

$z_{\alpha/2} = z_{0.025} = 1.96$ (from std normal table)

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

$\Rightarrow 0.29 < p < 0.39$ (95% CI).

Estimation of variances and chi-square distribution

Thm If $\bar{X}$ and s^2 are the sample mean and sample variance of a sample of size n from a normal population with mean μ and std deviation σ then:

$\frac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with $n-1$ degrees of freedom.

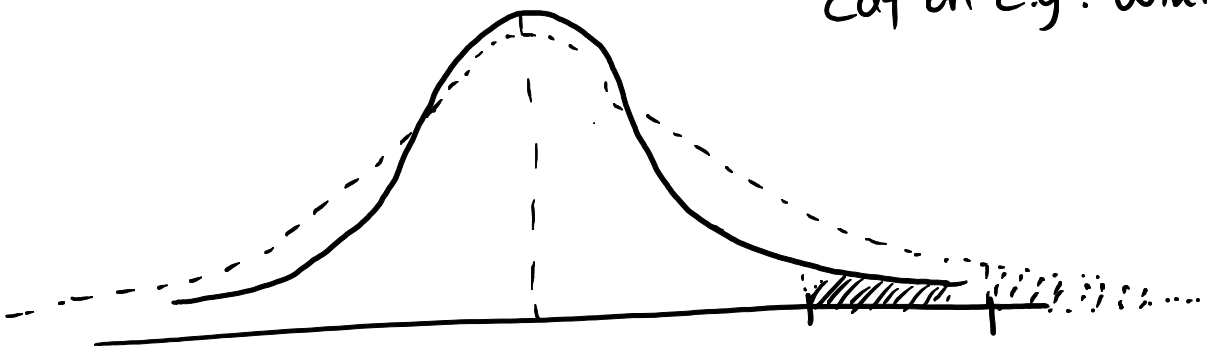
chi-square distribution

If $X_1, X_2, \ldots, X_n$ are independent standard normal random variables ($X_i \sim N(0,1)$ iid.) then $Y = \sum_{i=1}^{n} X_i^2$ has a chi-square distribution of degree n.

$\mu(Y) = n$ and $\sigma^2(Y) = 2n$.

Comparing normal and t-distributions

(see better plot of cdf on e.g. wikipedia)



$z_{0.025} = 1.96$ $\quad t_{0.025,4} = 2.776$

As $df \to \infty$, $t_{\alpha', df} \to z_{\alpha'}$ $(\alpha' = 1-\alpha)$

E.g. $t_{.025,4} = 2.776$

$t_{.025,15} = 2.131$

$\vdots$

$t_{.025,300} \approx 1.96$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{with} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

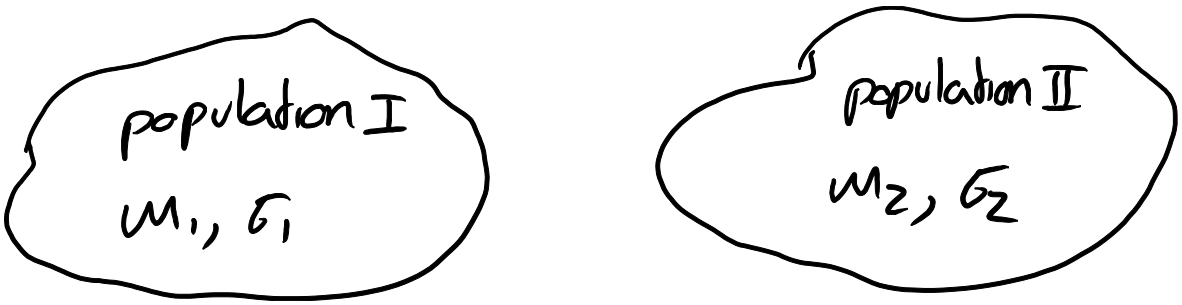
$$\sim \underbrace{S_t(0, 1, n-1)}$$

"standardized" t-distribution with $n-1$ degrees of freedom

Inferences about two population means

Large sample confidence interval

$n_1 \geq 30$, $n_2 \geq 30$, σ_1, σ_2 known



$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{x}_1 - x_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ex) Random samples of US and Japanese retail stores (same time period, uniqlo clothing stores; same goods; independent samples).

use $\alpha = 0.05$ 95% CI

US

$n_1 = 50$

$\bar{x}_1 = 11{,}545$ USD

$s_1 = 1989$ USD

Japan

$n_2 = 30$

$\bar{x}_2 = 12{,}243$ USD

$s_2 = 1{,}843$ USD

95% CI for $(\mu_1 - \mu_2)$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Here it is OK to substitute s_1^2 for σ_1^2 and s_2^2 for σ_2^2 as $n_1, n_2 \geq 30$ so samples are relatively large.

$$\Rightarrow (11545 - 12243) \pm \underbrace{1.96}_{z_{.025}} \sqrt{\frac{(1989)^2}{50} + \frac{(1843)^2}{30}}$$

$$\approx -698 \pm 859.60$$

Thus, we estimate the difference in retail sales to fall in the interval

$$-1557.60 \text{ to } 161.60$$

In other words, we estimate that μ_2 (the mean retail sales in Japan) could be larger than μ_1 (mean retail sale in US) by as much as 1557.60 or it could be

less than μ_1 by as much as 161.60.

small scale estimation for two populations:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with $s_p = \frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$

Can be shown that t-statistic has student t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

Ex) suppose TV network samples give the following viewer #s for broadcasts

sports	movies	$\alpha = 0.05$
$n_1 = 13$	$n_2 = 15$	95% CI
$\bar{x}_1 = 6.8$ mil	$\bar{x}_2 = 5.3$ mil	
$s_1 = 1.8$ mil	$s_2 = 1.6$ mil	

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{(13-1)(1.8)^2 + (15-1)(1.6)^2}{13+15-2} \approx 2.87$$

small sample CI for $(\mu_1 - \mu_2)$:

$$\underline{CI}\left\{ \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,\, n_1+n_2-2} \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \right\}$$

$$(6.8-5.3) \pm \underbrace{t_{\alpha/2,\, n_1+n_2-2}\sqrt{2.87\left(\frac{1}{13}+\frac{1}{15}\right)}}_{E}$$

$$t_{0.025,\,13+15-2} = t_{0.025,\,26} = 2.056$$

(from table)

$$\Rightarrow 2.056\sqrt{2.87\left(\frac{1}{13}+\frac{1}{15}\right)} = E \approx 1.32$$

$$\Rightarrow (\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

$$\Rightarrow 0.18 < \mu_1 - \mu_2 < 2.82 \quad \text{million people}$$

With about 95% confidence we can claim that more people watch sports than movies on this network.

Review

(I) Large scale estimation for mean

($n \geq 30$, σ known or $\sigma \approx s$):

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

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

$\Rightarrow P(0 < z < z_{\alpha/2}) = \frac{1-\alpha}{2}$

Ex) Find 95% CI for μ of a population which has variance $\sigma^2 = 100$. Consider sample size $n = 35$ with mean 67.53.

note: $n \geq 30$, CLT applies

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

sample means normally distributed

If $n < 30$, need population to be normal

$\alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$

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

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{10}{\sqrt{35}} \approx 3.31$$

$$\bar{X} - E < \mu < \bar{X} + E$$

$$\Rightarrow 64.21 < \mu < 70.84$$

(II) Small scale estimation for mean

Used when $n < 30$, population not known to be normal and when σ is unknown.

(Note, when $n \geq 30$, one can use $\sigma \approx s$, for smaller samples such estimation is inaccurate).

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \underbrace{S_t(0,1,n-1)}_{\text{standard student-t distrib. with } n-1 \text{ deg. of freedom}}$$

As $n \to \infty$, $S_t(0,1,n-1) \to N(0,1)$.

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} \quad \text{(sample variance)}$$

Ex) The nicotine contents of 5 cigarettes of a certain brand, measured in mg, are $\{21, 19, 23, 19, 23\}$. Establish a 99% CI of the avg nicotine content per cigarette of this brand.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{105}{5} = 21$$

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

x	$x-\bar{x}$	$(x-\bar{x})^2$
21	0	0
19	-2	4
23	2	4
19	-2	4
23	2	4

$$S = \sqrt{\frac{16}{5-1}} = 2 \text{ mg}$$

$$t_{0.005,\,4} = 4.604$$

$$E = t_{\alpha/2,\,n-1}\frac{s}{\sqrt{n}} = 4.604\,\frac{2}{\sqrt{4}} = 4.604$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$\Rightarrow 16.4 < \mu < 25.6$$

Notice that the interval is rather large for the given α ($\alpha = 0.01 \Rightarrow .99$ interval). This is because sample size is very small.

(III) proportions — when X is Binomial: $n \geq 50$ (large n).

$$\frac{X - np}{\sqrt{npq}} \to N(0,1) \quad \text{as } n \to \infty$$

(approximation of Binomial r.v. by normal)

For moderate n, approx is good when p not too small (not rare events).

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

$$\frac{X-np}{\sqrt{npq}} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} = \frac{\frac{X}{n}-p}{\sqrt{pq/n}} = \frac{\hat{p}-p}{\sqrt{pq/n}}$$

Use estimation $\sqrt{pq/n} \approx \sqrt{\hat{p}\hat{q}/n}$

$= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ valid by Law of Large numbers.

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

$$\Rightarrow \hat{p} - E < p < \hat{p} + E$$

where $\hat{p} = \frac{X}{n}$ (sample proportion)

Ex) 144 school children asked if or not they like windows in their classroom. 43 people preferred windows. Establish a .95 CI for proportion of elementary school children who like windows in classrooms.

$$\Rightarrow \quad \hat{p} = \frac{43}{144} \approx 0.30 \quad \text{(sample proportion)}$$

$n = 144$ large, $\hat{p}$ not too small

X : binomial r.v. counts # of children who like windows

$$\frac{X - np}{\sqrt{npq}} = \frac{\frac{X}{n} - p}{\sqrt{pq/n}} \sim N(0,1)$$

$$\sim \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{.30(.70)}{144}}$$

$$\approx 0.038 \times 1.96$$

$$\simeq 0.074$$

$$\hat{p} - E < p < \hat{p} + E$$

$$\Rightarrow \{.23 < p < .37\}$$ 95% CI for p