UCLA STAT 110A
Applied Statistics

- **Instructor:** Ivo Dinov,
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistant:** Helen Hu, UCLA Statistics

University of California, Los Angeles, Spring 2002

http://www.stat.ucla.edu/~dinov/

Inference & Estimation

- C + E model
- Types of Inference
- Sampling distributions
- CI's for μ & p
- Comparing 2 proportions
- How big should my study be?
- Paired vs. unpaired tests

The C + E Model

- **Data = Center + Error :** $Y = \mu + \varepsilon$;
- The response value **Y** is equal to unknown constant (μ), but because of normal variability we almost never observe μ exactly.
- Example **Speed of light** (SOL), $\mu = 2.998 \times 10^9$ *m/s*. However, 100 measurements of the SOL are all going to be slightly different.
- Model (population) parameter – a quantity describing the model that can take on many values. Ex., μ.

Types of inference

- **Estimation of model parameters:** Data-driven estimates of the model parameters. Also, includes how much uncertainty about those estimates is there.
- **Prediction of new (future) observations:** Uses past and current data to predict the value of new observations from the population.
- **Tolerance level:** a range of values that has user-specified probability of containing a particular proportion of the population.

Estimation of model parameter(s) – μ

- Least-Absolute-Error-Estimate(m) – Suppose, $\mu = 3.5$ (unknown) and $Y=\{Y_1=\mu+e_1, Y_2=\mu+e_2, \ldots, Y_{10}=\mu+e_{10}\}$ are our observed data. Cost function = Sum-of-Absolute-Errors = $SAE = \Sigma|Y_k - m| \rightarrow m = \text{MinArg}(SAE)$.
- Least-Squares(m) (in the same setting). Cost function = Sum-of-Squared-Errors = $SSE = \Sigma(Y_k - m)^2 \rightarrow$
  $m = \text{MinArg}(SSE)$, least-squares-estimate.
- Solution (differentiate):

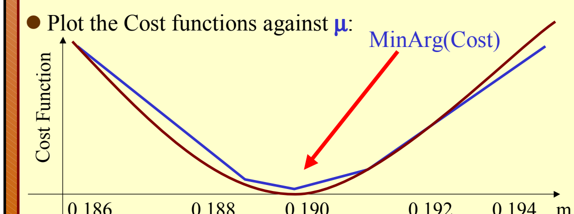
$$d\,SSE(m) / d\,m = -2\,\Sigma(Y_k - m) = 0, \text{ solve for m!}$$

Estimation of model parameter(s) – μ (Example)

- Data: ball-bearing diameter: μ =? (unknown) given the observed $\mathbf{Y=\{Y_1=0.1896, Y_2=0.1913, Y_{10}=0.1900\}}$.

$$\mathbf{SAE = \Sigma|Y_k - m| \quad \& \quad SSE = \Sigma(Y_k - m)^2}$$

- Plot the Cost functions against μ:



Parameters, Estimators, Estimates …

- A parameter is a characteristic of the data – mean, 1st quartile, SD, etc.)
- An estimator is an abstract **rule** for calculating a quantity (or parameter) from the sample data.
- An estimate is the value obtained when real data are plugged-in the estimator rule.

Parameters, Estimators, Estimates …

- E.g., We are interested in the population mean diameter (parameter) of washers the sample-average formula represents an estimator we can use, where as the value of the sample average for a particular dataset is the estimate (for the *mean* parameter).

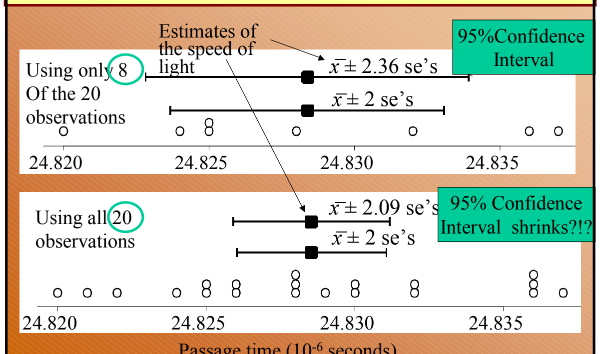
$$\text{parameter} = \mu_Y; \qquad \text{estimator} = \overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$$

$$\text{Data}: \ Y = \{0.1896, \ 0.1913, \ 0.1900\}$$

$$\text{estimate} = \overline{y} = \tfrac{1}{3}(0.1896 + 0.1913 + 0.1900)$$

$$\overline{y} = 0.1903. \ \text{How about } \overline{y} = \tfrac{2}{3}(0.1896 + 0.1913 + 0.1900)$$

20 replicated measurements to estimate the speed of light. Obtained by Simon Newcomb in 1882, by using distant (3.721 km) rotating mirrors.



A 95% confidence interval

- A type of interval that contains the true value of a parameter for 95% of samples taken is called a ***95% confidence interval*** for that parameter, the ends of the CI are called *confidence limits*.
- (For the situations we deal with) a confidence interval (CI) for the true value of a **parameter** is given by

estimate $\pm$ ***t standard errors (SE)***

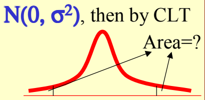
Value of the Multiplier, *t*, for a 95% CI

df :	7	8	9	10	11	12	13	14	15	16	17
t :	2.365	2.306	2.262	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110
df :	18	19	20	25	30	35	40	45	50	60	∞
t :	2.101	2.093	2.086	2.060	2.042	2.030	2.021	2.014	2.009	2.000	1.960

(General) Confidence Interval (CI)

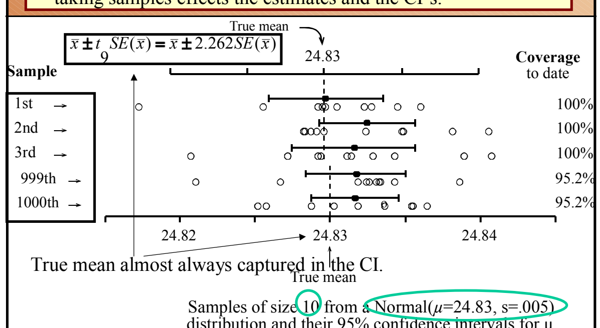
- A level L confidence interval for a parameter (θ), is an interval (θ_1^ , θ_2^), where θ_1^ & θ_2^, are estimators of θ, such that $P(\theta_1\text{^} < \theta < \theta_2\text{^}) = L$.
- E.g., **C+E model**: $Y = \mu + \varepsilon$. Where $\varepsilon \sim N(0, \sigma^2)$, then by CLT we have $Y\_bar \sim N(\mu, \sigma^2/n)$

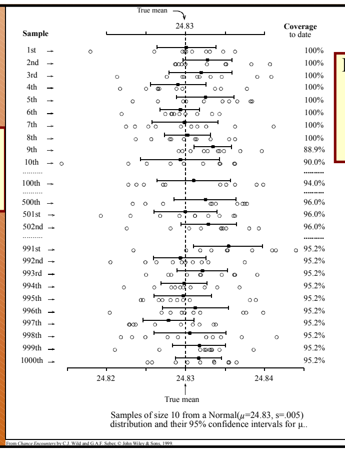
  ➔ $n^{1/2}(Y\_bar - \mu)/\sigma \sim N(0, \sigma^2)$.



- $L = P\left(z_{(1-L)/2} < n^{1/2}(Y\_bar - \mu)/\sigma < z_{(1+L)/2} \right)$, **where z_q is the q^{th} quartile.**
- **E.g., $0.95 = P\left(z_{0.025} < n^{1/2}(Y\_bar - \mu)/\sigma < z_{0.975} \right)$,**

- CI are constructed using the sample $\bar{x}$ and s=SE. But different samples yield different estimates and ➔ diff. CI's?!?
- Below is a computer simulation showing how the process of taking samples effects the estimates and the CI's.





Most of the table

How many of the previous samples contained the true mean?

CI for population mean

Confidence Interval for the true (population) mean μ:

sample mean* $\pm$ *t standard errors

or $\bar{x} \pm t \, \mathrm{se}(\bar{x})$, where $\mathrm{SE}(\bar{x}) = \frac{s_x}{\sqrt{n}}$ and $df = n - 1$

Value of the Multiplier, t, for a 95% CI

df :	7	8	9	10	11	12	13	14	15	16	17
t :	2.365	2.306	2.262	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110
df :	18	19	20	25	30	35	40	45	50	60	∞
t :	2.101	2.093	2.086	2.060	2.042	2.030	2.021	2.014	2.009	2.000	1.960

CI for population mean

- E.g., SYSTAT → Data: **BirthdayDistribution_1978_systat.SYD**
- Statistics → Descriptive Statistics → Stem-&-Leaf-Plot
- Statistics → Descriptive Statistics →CI_for_mean

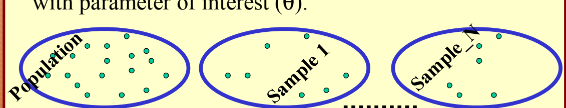
CI for population mean - Example

- E.g., Lab rats blood glucose levels:**{266, 149, 161, 220}** Estimate μ, the mean population blood sugar level. Assume the variance σ^2 =2958, → σ=54.4, from prior experience. Also assume data comes from $N(\mu, \sigma^2)$. Sample-avg=199, Compute the 95% CI, L=0.95.
- **(1-L)/2 = 0.025, (1+L)/2 = 0.975,**
- $\mathbf{Z_{(1-L)/2} = Z_{0.025} = -1.96 \quad \& \; Z_{(1+L)/2} = Z_{0.975} = 1.96}$
- $\mathbf{L = P\,(\, z_{(1-L)/2} \; < \; n^{1/2}(Y\_bar - \mu)/\sigma \; < \; z_{(1+L)/2}\,),}$
- $\mathbf{CI(\mu)= (Y\_bar - \sigma z_{(1+L)/2}/n^{1/2} \; ; Y\_bar - \sigma z_{(1-L)/2}/n^{1/2})}$
- $\mathbf{CI(\mu)= (199 - 54.4x1.96 / 4^{1/2} \; ; 199 + 54.4x1.96 / 4^{1/2} \,)}$

  $\mathbf{CI(\mu)= (\,145.7 \; : \; 252.3\,)}$

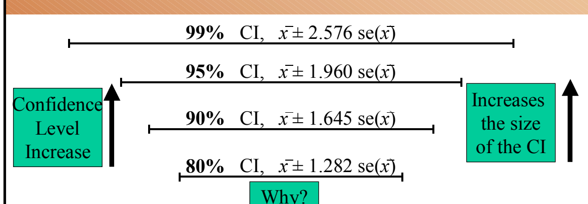
CI - Interpretation

- Consider taking all possible samples from the population with parameter of interest (θ).



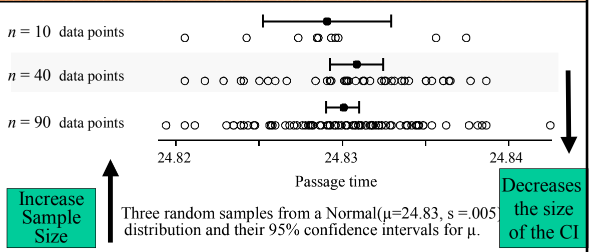
- Suppose we construct the level L confidence interval for a parameter (θ) for each sample. Then a proportion L of all constructed CI's will contain the value of θ.
- Note that this interpretation of CI's is in terms of **repeated sampling** from the same population …

Effect of increasing the confidence level



The greater the confidence level, the wider the interval

Effect of increasing the sample size



Three random samples from a Normal(μ=24.83, s =.005) distribution and their 95% confidence intervals for μ.

To ***double the precision*** we need ***four times*** as many observations.

Why ↑ in sample-size ↓ CI?

Confidence Interval for the true (population) mean μ:

sample mean $\pm$ ***t standard errors***

or $\quad \bar{x} \pm t\ \mathrm{se}(\bar{x})$, where $\mathrm{se}(\bar{x}) = \dfrac{s_x}{\sqrt{n}}$ and $df = n-1$

Student's *t*-distribution

- For random samples from a Normal distribution,

$$T = \frac{(\bar{X} - \mu)}{SE(\bar{X})}$$

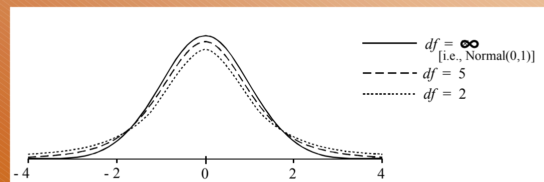
Recall that for samples from N(μ , σ)

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

is exactly distributed as Student($df = n$ - 1) ← Approx/Exact Distributions

- but methods we shall base upon this distribution for T work well even for small samples sampled from distributions which are quite non-Normal.
- df is number of observations –1, degrees of freedom.

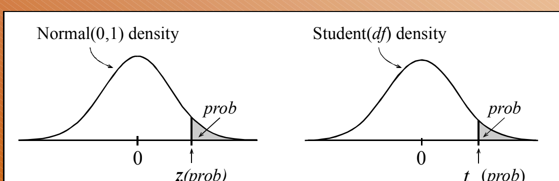
Density curves for Student's *t*



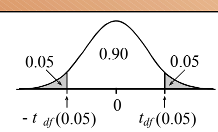
Student(df) density curves for various *df*.

Notation

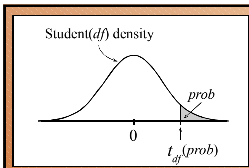
- By $t_{df}(prob)$, we mean the number t such that when $T \sim$ Student(df), P($T \geq t_{df}$) = $prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.



The *z(prob)* and *t(prob)* notations.



The central 90% of the Student(*df*) distribution.



Reading Student's *t* table

Desired *upper-tail prob*

Extracts from the Student's t-Distribution Table

prob

Desired *df*

df	.20	.15	.10	.05	.025	.01	.005	.001	.0005	.0001
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
…	…	…	…	…	…	…	…	…	…	…
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587	5.694
…	…	…	…	…	…	…	…	…	…	…
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
…	…	…	…	…	…	…	…	…	…	…
∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

t-value

Comparison of the CI using T (unknown σ) & Z (known σ) distributions

- **For the old data:** glucose levels **:** **{266, 149, 161, 220}** $\hat{\sigma}=\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(y_k-\bar{y})^2}$
- **CI(μ), when σ is unknown** (T-distr.), small-sample-size, and data comes from (approx.) Normal distribution.

$$\bar{x}=199$$
$$\hat{\sigma}=54.39$$

L= P ($t_{N-1,(1-L)/2}$< $n^{½}$(Y_{bar}- μ)/σ^ <$t_{N-1,(1+L)/2}$) ,

CI(μ)=(Y_{bar}– σ^$t_{N-1,(1+L)/2}$/$n^{½}$; Y_{bar}- σ^$t_{N-1,(1+L)/2}$/$n^{½}$)

95% CI(μ)=(199–54.39x3.18 /4½; 199+54.39x3.18 / 4½)

$t_{N-1,(1+L)/2}$ = $t_{3,0.975}$=3.18 & $t_{N-1,(1-L)/2}$ = $t_{3,0.025}$=-3.18 → CI_T(μ)=(112.4:285.6)

Comparison of the CI using T (unknown σ) & Z (known σ) distributions

- **CI(μ), when σ = 54.4 is known** (Normal distr.)

CI(μ) = (Y_{bar}– σ $z_{(1+L)/2}$/$n^{½}$;Y_{bar}- σ $z_{(1+L)/2}$/$n^{½}$),

$z_{(1+L)/2}$ = 1.96

95% CI(μ) = (199 – 54.4 x 1.96 /4½ ; 199+54.4 x 1.96 / 4½)

CI_Z(μ)= (145.7 : 252.3)

- **Comparison:**

CI_T(μ)=(112.4:285.6) ← compare→

CI_Z(μ)= (145.7:252.3)

Which one is better?!? More appropriate?!?

Prediction vs. Confidence intervals

- **Confidence Intervals (for the population mean μ):**

$$\left(\bar{Y}-\frac{\hat{\sigma}\times t_{n-1,(1+L)/2}}{\sqrt{n}}\ ;\ \bar{Y}+\frac{\hat{\sigma}\times t_{n-1,(1+L)/2}}{\sqrt{n}}\right)$$

- **Prediction Intervals:** L-level prediction interval (**PI**) for a new value of the process Y is defined by:

$$\left(\hat{Y}_{new}-\hat{\sigma}\times t_{n-1,(1+L)/2}\ ;\ \hat{Y}_{new}+\hat{\sigma}\times t_{n-1,(1+L)/2}\right)$$

where the predicted value $\hat{Y}_{new}=\bar{Y}$, is obtained as an estimator of the unknown process mean μ.

Prediction vs. Confidence intervals – Differences?

- **Confidence Intervals (for the population mean μ):**

$$\left(\bar{Y}-\frac{\hat{\sigma}\times t_{n-1,(1+L)/2}}{\sqrt{n}}\ ;\ \bar{Y}+\frac{\hat{\sigma}\times t_{n-1,(1+L)/2}}{\sqrt{n}}\right)$$

$$\hat{\sigma}=\hat{\sigma}(\bar{Y})=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2}$$

Which SD is bigger?!?

- **Prediction Intervals:**

$$\left(\hat{Y}_{new}-\hat{\sigma}\times t_{n-1,(1+L)/2}\ ;\ \hat{Y}_{new}+\hat{\sigma}\times t_{n-1,(1+L)/2}\right),\quad \text{where}\quad \hat{Y}_{new}=\bar{Y}$$

$$\hat{\sigma}=\hat{\sigma}(Y_{new}-\hat{Y}_{new})=\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2}\times\sqrt{1+\frac{1}{n}}$$

Classical Prediction for the C+E model

- **Y = C + E. When why, how to use prediction?**
- **When: E ~ *N*(0, σ²) ← → Y ~ *N*(μ, σ²),** there are more general situations, of course. Here we only consider this case.
- **Why:** Future predictions are of paramount importance in any area of science/engineering/medicine.
- **How: μ** is mostly unknown, so we estimate it by: **m^, (the sample average).**

If **population proportion, p,** is unknown we estimate it by the **sample-proportion, p^,** etc.

Classical Prediction for the C+E model

- **How:** μ is mostly unknown, so we estimate it by: **m^,**
  - Let $\hat{Y}_{new}$ be the predicted value
  - Error made by using $\hat{Y}_{new}$, instead of observing a new value, Y_{new} is:

    (1) $\mathbf{Y_{new} - \hat{Y}_{new} = (\mu - \varepsilon_{new}) - \hat{Y}_{new} = (\mu - \hat{Y}_{new}) + \varepsilon_{new}}$
  - But if we use $\hat{\mu}$ to predict a new value for Y, $\hat{Y}_{new} = \hat{\mu}$.
  - $Var(\mu - \hat{Y}_{new}) = Var(\hat{Y}_{new}) = Var(\hat{\mu}) = Var(\text{SampleAvg}) = \sigma^2/n$.
  - The variance of the second term is just σ^2.
  - Since the first-term in (1) is obtained from $\{Y_1, Y_2, \ldots, Y_n\}$, and $\varepsilon_{new} = \varepsilon_{n+1}$, we have two independent terms ➔ Variances add up!
  - $Var(\mathbf{Y_{new} - \hat{Y}_{new}}) = Var(\mu - \hat{Y}_{new}) + Var(\varepsilon_{new}) = \sigma^2/n + \sigma^2$.

Classical Prediction for the C+E model

- **How:** Let $\hat{Y}_{new}$ be the predicted value
  - Error $\mathbf{Y_{new} - \hat{Y}_{new} = (\mu - \varepsilon_{new}) - \hat{Y}_{new} = (\mu - \hat{Y}_{new}) + \varepsilon_{new}}$
  - $Var(\mathbf{Y_{new} - \hat{Y}_{new}}) = Var(\mu - \hat{Y}_{new}) + Var(\varepsilon_{new}) = \sigma^2/n + \sigma^2$.
  - Often σ is unknown, and we estimate it by the sample SD, S ➔
  - $SD(\mathbf{Y_{new} - \hat{Y}_{new}}) = [S^2(\mathbf{1+1/n})]^{1/2}$
- We can show that $$T = \frac{Y_{new} - \hat{Y}_{new} - 0}{\sigma(Y_{new} - \hat{Y}_{new})} \sim t_{n-1}$$
- ➔ The L-level prediction interval (**PI(Y$_{new}$)**) is:

$$L = P(t_{n-1,(1-L)/2} < T < t_{n-1,(1+L)/2}) \rightarrow$$
$$\left(\hat{Y}_{new} + \hat{\sigma} \times t_{n-1,(1-L)/2} \;;\; \hat{Y}_{new} + \hat{\sigma} \times t_{n-1,(1+L)/2}\right)$$
$$\left(\hat{Y}_{new} - \hat{\sigma} \times t_{n-1,(1+L)/2} \;;\; \hat{Y}_{new} + \hat{\sigma} \times t_{n-1,(1+L)/2}\right)$$

Solve for T
By symmetry of t_{n-1}.

CI for a population proportion

Confidence Interval for the true (population) proportion p:

sample proportion ± z standard errors

or $\hat{p} \pm z\ \text{se}(\hat{p})$, where $\text{se}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

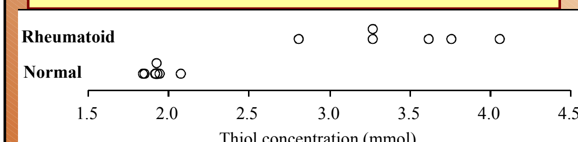
Example – higher blood thiol concentrations associated with rheumatoid arthritis?!?

Thiol Concentration (mmol)

	Normal	Rheumatoid
	1.84	2.81
	1.92	4.06
	1.94	3.62
	1.92	3.27
	1.85	3.27
	1.91	3.76
	2.07	
Sample size	7	6
Sample mean	1.92143	3.46500
Sample standard deviation	0.07559	0.44049

Research question: Is the change in the Thiol status in the lysate of packed blood cells substantial to be indicative of a non trivial relationship between Thiol-levels and rheumatoid arthritis?

Example – higher blood thiol concentrations with rheumatoid arthritis



Dot plot of Thiol concentration data.

Two groups of subjects are studied: 1. NC (normal controls)
2. RA (rheumatoid arthritis).

Observations: 1. The avg. levels of thiol seem diff. in NC & RA
2. NC and RA groups are separated completely.

Question: Is there statistical evidence that thiol-level correlates with the disease?

Difference between means

Confidence Interval for a difference between population means $(\mu_1 - \mu_2)$:

Difference between sample means ± t standard errors of the difference

or $\boxed{\bar{x}_1 - \bar{x}_2 \pm t\ \text{se}(\bar{x}_1 - \bar{x}_2)}$

Difference between proportions

Confidence Interval for a difference between population proportions $(p_1 - p_2)$:

Difference between sample proportions
± z standard errors of the difference

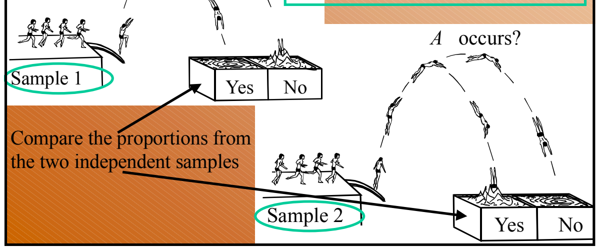
$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z\,\mathrm{se}(\hat{p}_1 - \hat{p}_2)}$$

Big Question ???

How do we compute the SE($\hat{p}_1 - \hat{p}_2$) for different cases?

Proportions from 2 independent samples

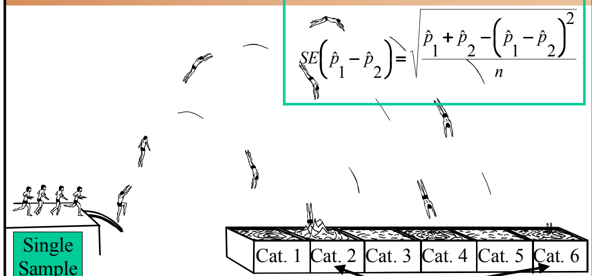
$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_1}}$$



Compare the proportions from the two independent samples

Single sample, several response categories

$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - \left(\hat{p}_1 - \hat{p}_2\right)^2}{n}}$$



Compare different proportions from the same sample

Example – 1996 US Presidential Election

		Pre-election Polls				Election Results		
State	n	Clinton	Doll	Perot	Other/Undecided	Clinton	Doll	Perot
New Jersey	1,000	51	33	8	8	53	36	9
New York	1,000	59	25	7	9	59	31	8
Connecticutt	1,000	51	29	11	9	52	35	10

Compare proportions of NJ and NY voters supporting Clinton and Dole, pre- and post election

$$\hat{p}_1 - \hat{p}_2 \pm z\,\mathrm{se}(\hat{p}_1 - \hat{p}_2)$$

Note the independence-case SE formula is only applicable for the cases when the samples are independent. In this case, the pre-election poll and the election results are **not independent** (obviously these are highly correlated observations).

Example – 1996 US Presidential Election

		Pre-election Polls				Election Results		
State	n	Clinton	Doll	Perot	Other/Undecided	Clinton	Doll	Perot
New Jersey	1,000	51	33	8	8	53	36	9
New York	1,000	59	25	7	9	59	31	8
Connecticutt	1,000	51	29	11	9	52	35	10

Proportions from 2 independent samples

How far is Clinton ahead In NY Compared to NJ?
Diff.proportions= 59-51%=8%
CI: [4% : 12%]
Actual diff 59-53=6

$$\hat{p}_1 - \hat{p}_2 \pm z\,\mathrm{se}(\hat{p}_1 - \hat{p}_2)$$

$$\text{estimate} \pm z \times \mathrm{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$0.08 \pm 1.96 \times 0.02842 = [4\%; 12\%]$$

Example – 1996 US Presidential Election

		Pre-election Polls				Election Results		
State	n	Clinton	Doll	Perot	Other/Undecided	Clinton	Doll	Perot
New Jersey	1,000	51	33	8	8	53	36	9
New York	1,000	59	25	7	9	59	31	8
Connecticutt	1,000	51	29	11	9	52	35	10

Single sample, several response categories

How far is Clinton ahead of Dole in NJ?
Diff.proportions= 18%
CI: [12% : 24%]
Actual diff 53-36=17

$$\hat{p}_1 - \hat{p}_2 \pm z\,\mathrm{se}(\hat{p}_1 - \hat{p}_2)$$

$$\text{estimate} \pm z \times \mathrm{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - \left(\hat{p}_1 - \hat{p}_2\right)^2}{n}} =$$

$$0.18 \pm 1.96 \times 0.02842 = [12\% : 24\%]$$

SE's for the 2 cases of differences in proportion

(a) **Proportions from two independent samples** of sizes n_1 and n_2, respectively

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

(b) **One sample of size n, several response categories**

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

Sample size - proportion

- **For a 95% CI,** $\text{margin} = 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$
- **Sample size for a desired margin of error:** For a margin of error no greater than *m*, use a sample size of approximately

$$n = \left(\frac{z}{m}\right)^2 \times p^*(1-p^*)$$

- p^* is a guess at the value of the proportion -- err on the side of being too close to 0.5
- *z* is the multiplier appropriate for the confidence level
- *m* is expressed as a proportion (between 0 and 1), not a percentage (basically, What's n, so that **m >= margin**?)

Sample size -- mean

- **Sample size for a desired margin of error:** For a margin of error no greater than *m*, use a sample size of approximately

$$n = \left(\frac{z\sigma^*}{m}\right)^2$$

- σ^* is an estimate of the variability of individual observations
- *z* is the multiplier appropriate for the confidence level

Paired vs. Unpaired comparisons

- We will discuss these later, when we get to the hypothesis testing (ch6_HT_Paired_Indep_Tests.ppt)

Confidence intervals

- We construct an interval estimate of a parameter to summarize our level of uncertainty about its true value.
- The uncertainty is a consequence of the sampling variation in point estimates.
- If we use a method that produces intervals which contain the true value of a parameter for 95% of samples taken, the interval we have calculated from our data is called a 95% confidence interval for the parameter.
- Our confidence in the particular interval comes from the fact that the method works 95% of the time (for 95% CI's).

Summary cont.

- For a great many situations,

an (approximate) confidence interval is given by

estimate $\pm$ ***t standard errors***

The size of the multiplier, *t*, depends both on the desired confidence level and the degrees of freedom (*df*).

[With proportions, we use the Normal distribution (i.e., $df=\infty$) and it is conventional to use *z* rather than *t* to denote the multiplier.]

- The *margin of error* is the quantity added to and subtracted from the estimate to construct the interval (i.e. *t* standard errors).

Summary cont.

- If we want greater confidence that an interval calculated from our data will contain the true value, we have to use a wider interval.
- To double the precision of a 95% confidence interval (i.e.halve the width of the confidence interval), we need to take 4 times as many observations.