UCLA STAT 110 A
Applied Probability & Statistics for Engineers

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

University of California, Los Angeles, Spring 2003
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 = $SAE = \Sigma|Y_k - m| \rightarrow m = MinArg(SAE)$.
- Least-Squares(m) (in the same setting). Cost function = Sum-of-Squared-Errors = $SSE = \Sigma(Y_k - m)^2 \rightarrow$

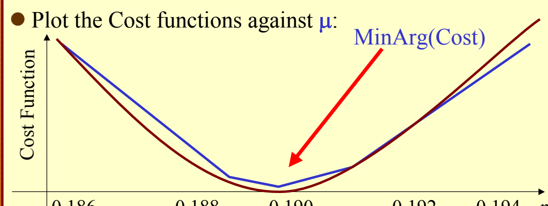
  $m = MinArg(SSE)$, least-squares-estimate.
- Solution (differentiate):

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

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

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

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

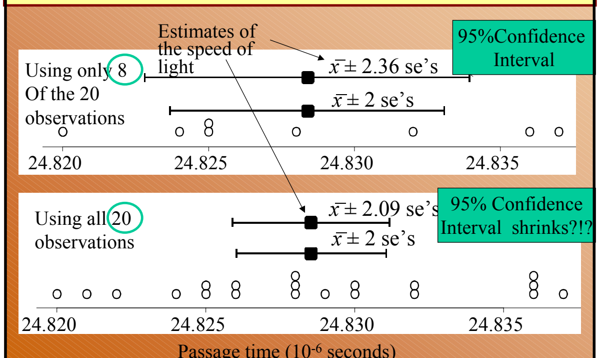
$$\text{parameter} = \mu_Y; \qquad \text{estimator} = \overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$$

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

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

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

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)***

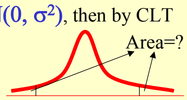
Value of the Multiplier, *t*, for a 95% CI

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)

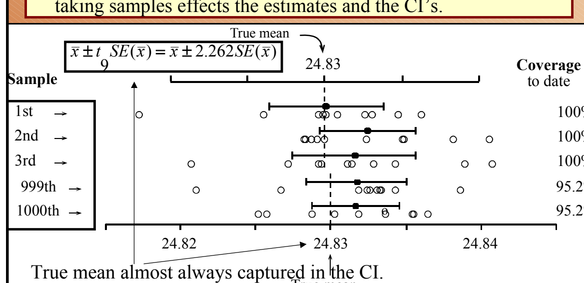
- A **level L confidence interval** for a parameter (θ), is an interval ($\hat{\theta}_1$, $\hat{\theta}_2$), where $\hat{\theta}_1$ & $\hat{\theta}_2$, are estimators of θ, such that $P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = L$.
- E.g., **C+E model**: $Y = \mu + \varepsilon$. Where $\varepsilon \sim N(0, \sigma^2)$, then by CLT we have $\bar{Y} \sim N(\mu, \sigma^2/n)$

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



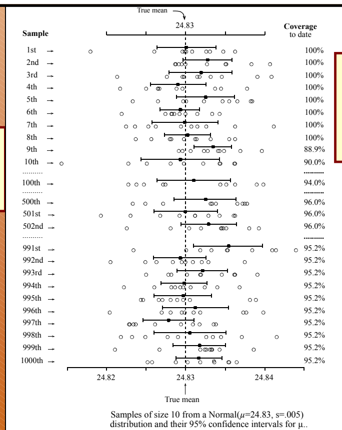
- $L = P\left(z_{(1-L)/2} < n^{1/2}(\bar{Y} - \mu)/\sigma < z_{(1+L)/2} \right)$, **where z_q is the q^{th} quartile.**
- **E.g., $0.95 = P\left(z_{0.025} < n^{1/2}(\bar{Y} - \mu)/\sigma < z_{0.975} \right)$,**

- 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.



Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ..

Most of the table



Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ.

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

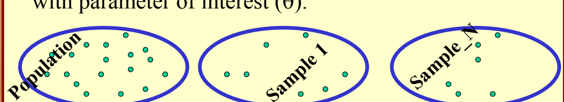
CI for population mean - Example

- 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

99% CI, $\bar{x} \pm 2.576\ \mathrm{se}(\bar{x})$

95% CI, $\bar{x} \pm 1.960\ \mathrm{se}(\bar{x})$

Confidence Level Increase

90% CI, $\bar{x} \pm 1.645\ \mathrm{se}(\bar{x})$

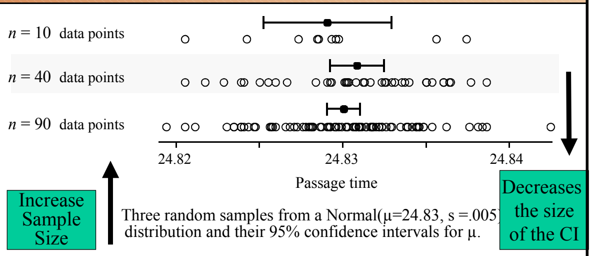
Increases the size of the CI

80% CI, $\bar{x} \pm 1.282\ \mathrm{se}(\bar{x})$

Why?

The greater the confidence level, the wider the interval

Effect of increasing the sample size



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

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

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 ($t_{N-1, (1-L)/2}$ < $n^{½}(Y_{bar} - \mu)/\sigma^\wedge$ < $t_{N-1, (1+L)/2}$) ,

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

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) = ($Y_{bar} - \sigma\ z_{(1+L)/2}/n^{½}$;$Y_{bar} - \sigma\ z_{(1+L)/2}/n^{½}$),

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

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

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

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

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):**

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

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

Which SD is bigger?!?

- **Prediction Intervals:**

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

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

Classical Prediction for the C+E model

- $Y = C + E$. **When why, how to use prediction?**
- **When: E ~ $N(0, \sigma^2)$ ⇐ ⇒ Y ~ $N(\mu, \sigma^2)$,** 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 $Y^\wedge_{new}$ be the predicted value
  - Error made by using $Y^\wedge_{new}$, instead of observing a new value, Y_{new} is:

    (1) $\mathbf{Y_{new} - Y^\wedge_{new} = (\mu - \varepsilon_{new}) - Y^\wedge_{new} = (\mu - Y^\wedge_{new}) + \varepsilon_{new}}$
  - But if we use $\mu^\wedge$ to predict a new value for Y, $Y^\wedge_{new} = \mu^\wedge$.
  - $Var(\mu - Y^\wedge_{new}) = Var(Y^\wedge_{new}) = Var(\mu^\wedge) = Var(\text{SampleAvg}) = \sigma^2/n$.
  - 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!
  - $Var(\mathbf{Y_{new} - Y^\wedge_{new}}) = Var(\mu - \mathbf{Y^\wedge_{new}}) + Var(\varepsilon_{new}) = \sigma^2/n + \sigma^2$.

Classical Prediction for the C+E model

- **How:** Let $Y^\wedge_{new}$ be the predicted value
  - Error $\mathbf{Y_{new} - Y^\wedge_{new} = (\mu - \varepsilon_{new}) - Y^\wedge_{new} = (\mu - Y^\wedge_{new}) + \varepsilon_{new}}$
  - $Var(\mathbf{Y_{new} - Y^\wedge_{new}}) = Var(\mu - \mathbf{Y^\wedge_{new}}) + Var(\varepsilon_{new}) = \sigma^2/n + \sigma^2$.
  - Often σ is unknown, and we estimate it by the sample SD, S ➔
  - SD $(\mathbf{Y_{new} - Y^\wedge_{new}}) = [S^2(\mathbf{1+1/n})]^{1/2}$
- We can show that

$$T = \frac{Y_{new} - \hat{Y}_{new} - 0}{\sigma\left(Y_{new} - \hat{Y}_{new}\right)} \sim t_{n-1}$$

- ➔ The L-level prediction interval (**PI($\mathbf{Y_{new}}$)**) 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 $\pm$ ***z standard errors***

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

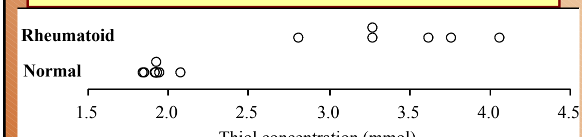
Example – higher blood thiol concentrations associated with rheumatoid arthritis?!?

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?

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

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

$\pm$ t standard errors of the difference

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

Example – higher blood thiol concentrations with rheumatoid arthritis

Confidence Interval for a difference between population means $(\mu_1 - \mu_2)$: $\boxed{\bar{x}_1 - \bar{x}_2 \pm t\ \mathrm{se}(\bar{x}_1 - \bar{x}_2)}$

or $$\bar{x}_1 - \bar{x}_2 \pm t\ \mathrm{se}(\bar{x}_1 - \bar{x}_2) = \\ 1.92 - 3.47 \pm t_{6-1,\,0.025}\sqrt{0.08^2 + 0.44^2} = \\ -1.55 \pm 2.571 \times 0.45 = \\ -1.55 \pm 1.15$$

Difference between proportions

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

Difference between sample proportions

$\pm$ z standard errors of the difference

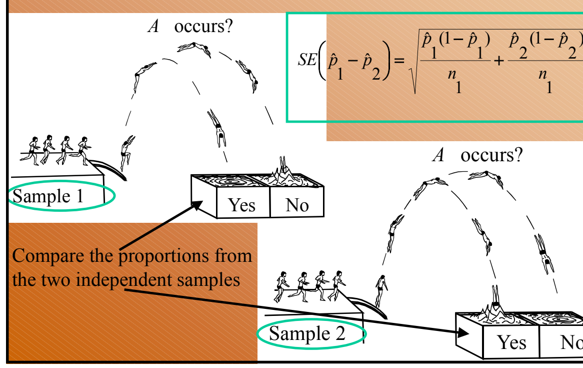
$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z\ \mathrm{se}(\hat{p}_1 - \hat{p}_2)}$$

Big Question ???

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

Proportions from 2 independent samples

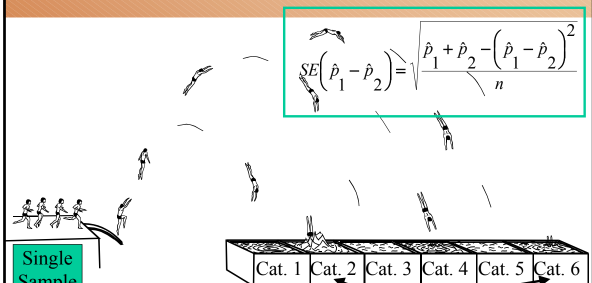
$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$



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

State	n	Pre-election Polls				Election Results		
		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

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

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

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

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

Example – 1996 US Presidential Election

State	n	Pre-election Polls				Election Results		
		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

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

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

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

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

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.

Examples – Birthday Paradox

- The Birthday Paradox: In a random group of N people, what is the change that at least two people have the same birthday?
- E.x., if N=23, P>0.5. Main confusion arises from the fact that in real life we rarely meet people having the same birthday as us, and we meet more than 23 people.
- The reason for such high probability is that any of the 23 people can compare their birthday with any other one, not just you comparing your birthday to anybody else's.
- There are N-Choose-2 = 20*19/2 ways to select a pair or people. Assume there are 365 days in a year, P(one-particular-pair-same-B-day)=1/365, and
- P(one-particular-pair-failure)=1-1/365 ~ 0.99726.
- For N=20, 20-Choose-2 = 190. E={No 2 people have the same birthday is the event all 190 pairs fail (have different birthdays)}, then $P(E) = P(\text{failure})^{190} = 0.99726^{190} = 0.59$.
- Hence, P(at-least-one-success)=1-0.59=0.41, quite high.
- Note: for N=42 ➔ P>0.9 …

Confidence intervals – non-symmetric case

- A marine biologist wishes to use male angelfish for an experiment and hopes their weights don't vary much. In fact, a previous random sample of n = 16 angelfish yielded the data below
- $\{y_1; \ldots ; y_n\}$ = { 5.1; 2.5; 2.8; 3.4; 6.3; 3.6; 3.9; 3.0; 2.7; 5.7; 3.5; 3.6; 5.3; 5.1; 3.5; 3.3}
- Sample statistics from these data include Avg. = 3.96 lbs, $s^2 = 1.35$ lbs, n = 16.
- **Problem**: Obtain a 100(1- α)% CI(σ^2).
- Point Estimator for σ^2? How about sample variance, s^2?
- Sampling theory for s^2? Not in general, but under Normal assumptions …
- If a random sample $\{Y_1; \ldots ;Y_n\}$ is taken from a normal population with mean μ and variance σ^2,then standardizing, we get a sum of squared N(0,1)

Confidence intervals – non-symmetric case

- $\{y_1; \ldots ; y_n\}$ = { 5.1; 2.5; 2.8; 3.4; 6.3; 3.6; 3.9; 3.0; 2.7; 5.7; 3.5; 3.6; 5.3; 5.1; 3.5; 3.3}
- **Problem**: Obtain a 100(1- α)% CI(σ^2).
- If a random sample $\{Y_1; \ldots ;Y_n\}$ is taken from a normal population with mean μ and variance σ^2,then standardizing, we get a sum of squared N(0,1)

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{\sigma^2} \sim \chi^2_{df=n-1}$$

For a=0.05, say. Need: 100(1- α)% CI(σ²).

$$\Rightarrow 1-\alpha = P\left(\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)} \le \frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{\sigma^2} \le \chi^2_{\left(n-1,\frac{\alpha}{2}\right)}\right)$$

Confidence intervals – non-symmetric case

- $\{y_1; \ldots ; y_n\}$ = { 5.1; 2.5; 2.8; 3.4; 6.3; 3.6; 3.9; 3.0; 2.7; 5.7; 3.5; 3.6; 5.3; 5.1; 3.5; 3.3}
- **Problem**: Obtain a 100(1- α)% CI(σ^2).

$$\frac{\sum_{k=1}^{n}\left(Y_k-\bar{Y}\right)^2}{\chi^2_{\left(n-1,\frac{\alpha}{2}\right)}} \le \sigma^2 \le \frac{\sum_{k=1}^{n}\left(Y_k-\bar{Y}\right)^2}{\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}}$$

- $\chi^2(15; 0.025)$=27:49 and $\chi^2(15; 0.975)$=6:26 ➔
- This yields the CI, the sample variance is **s^2=1.35**. Note the CI is NOT symmetric **(0.74 ; 3.24)**

Confidence intervals – non-symmetric case

- $\{y_1; \ldots ; y_n\}$ = { 5.1; 2.5; 2.8; 3.4; 6.3; 3.6; 3.9} & $\{x_1; \ldots ; x_k\}$={3.0; 2.7; 5.7; 3.5; 3.6; 5.3; 5.1; 3.5; 3.3}
- **Problem**: Obtain a 100(1- α)% CI(σ^2_y / σ^2_x). Diff variances?

$$\frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2 \Big/ (n-1)\times\sigma_Y^2}{\sum_{j=1}^{k}\left(X_j-\bar{X}\right)^2 \Big/ (k-1)\times\sigma_X^2} \sim F(n\text{-}1,\ k\text{-}1)$$

$$\Rightarrow 95\%\ CI\left(\sigma_Y^2 \big/ \sigma_X^2\right) = \ldots$$

- Ratio of two χ^2 variables is F-distributed …

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.

Parameter (Point) Estimation

- (6.2) Two Ways of Proposing Point Estimators
- **Method of Moments** (MOMs):
  - Set your k parameters equal to your first k moments.
  - Solve. (e.g., Binomial, Exponential and Normal)
- **Method of Maximum Likelihood** (MLEs):
- 1. Write out likelihood for sample of size n.
- 2. Take natural log of the likelihood.
- 3. Take partial derivatives with respect to your k parameters.
- 4. Take second derivatives to check that a maximum exists(f ”>0).
- 5. Set 1st derivatives equal to zero and solve for MLEs. e.g., Binomial, Exponential and Normal

Parameter (Point) Estimation

- Suppose we flip a coin n=8 times and observe {T,H,T,H,H,T,H,H}. Estimate the value p = P(H).
- **Method of Moments Estimate p^**:
  - Set your k parameters equal to your first k moments.
- Let X = {# T's} ➔ np=8p=E(X)= Sample#H's = 5 ➔ p^=5/8.
- **Method of Maximum Likelihood Estimate p^**:
- 1. f(x | p) = $\binom{8}{5}p^5(1-p)^3$ likelihood function.
- 2. $\ln\left(\binom{8}{5}p^5(1-p)^3\right) = \ln\left(\binom{8}{5}\right) + 5\times\ln(p) + 3\times\ln(1-p)$
- 3. $\frac{d\left(\ln\left(\binom{8}{5}\right) + 5\times\ln(p) + 3\times\ln(1-p)\right)}{d\,p} = \frac{5}{p} - \frac{3}{1-p} = 0$

$5(1-p) - 3p = 0 \Rightarrow p = 5/8$

Example – Maximum Likelihood Estimate

- Let $\{X_1, \ldots, X_n\}$={0.5, 0.3, 0.6, 0.1, 0.2}, weights, be IID N(μ, 1) ➔ f(x;μ). **Joint density** is $f(x_1,\ldots,x_n;\ \mu)=f(x_1;\mu)\times\ldots\times f(x_n;\mu)$.
- The likelihood function $L(p) = f(X_1,\ldots,X_n;\ p)$

$$L(\mu) = \lambda(x_1,\ldots,x_n) =$$

$$= e^{-\frac{(0.5-\mu)^2+(0.3-\mu)^2+(0.6-\mu)^2+(0.1-\mu)^2+(0.2-\mu)^2}{2}}$$

$$\ln(L) = (-1/2)\left[(0.5-\mu)^2+(0.3-\mu)^2+(0.6-\mu)^2+(0.1-\mu)^2+(0.2-\mu)^2\right]$$

$$0 = \frac{d\ln(L)}{d\mu} = (0.5-\mu)+(0.3-\mu)+(0.6-\mu)+(0.1-\mu)+(0.2-\mu) =$$

$$= -5\mu + 1.7 \Rightarrow \mu = 0.34 \Rightarrow \frac{d^2\ln(L)}{d\mu^2} = -5 \Rightarrow L(\mu = 0.34) = \max$$