UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: **Ivo Dinov,**

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Jacquelina Dacosta & Chris Barr

University of California, Los Angeles, Fall 2006

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

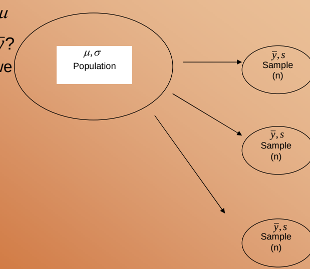
Sampling Distribution for the Mean and Introduction to Confidence Intervals

Quantitative Data

- More complex than dichotomous data
- Sample and populations for quantitative data can be described in various ways: mean, median, standard deviation (each has it's own sampling distribution.)

Sampling Distribution of $\bar{y}$

- Recall: $\bar{y}$ is used to estimate μ
- Question: How close to μ is $\bar{y}$?
  - Before we can answer this we need to define the probability distribution that describes sampling variability of $\bar{y}$



Sampling Distribution of $\bar{y}$

- Two really important facts:
  - The average of the sampling distribution of $\bar{y}$ is μ
    - Notation: $\mu_{\bar{y}} = \mu$
  - The standard deviation of the sampling distribution of $\bar{y}$ is $\frac{\sigma}{\sqrt{n}}$
    - Notation: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
  - Note: As n → ∞, $\sigma_{\bar{y}}$ gets smaller
  - Why? Look at the formula
  - Intuitively does this make sense?

Sampling Distribution of $\bar{y}$

- **Theorem 5:1** p.159
  - $\mu_{\bar{y}} = \mu$ (mean of the sampling distribution of $\bar{y}$ = μ the population mean)
  - $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (standard deviation (sd) of the sampling distribution of $\bar{y}$ = $\frac{\sigma}{\sqrt{n}}$ the population SD divided by $\sqrt{n}$)
  - Shape:
    - If the distribution of Y is normal the sampling distribution of $\bar{y}$ is normal.
    - Central Limit Theorem (CLT) - If n is large, then the sampling distribution of $\bar{y}$ is approximately normal, even if the population distribution of Y is not normal.

Central Limit Theorem (CLT)

- No matter what the distribution of Y is, if n is large enough the sampling distribution of $\bar{y}$ will be approximately normally distributed
  - HOW LARGE??? Rule of thumb $n \geq 30$.
- The closeness of $\bar{y}$ to μ depends on the sample size
- The more skewed the distribution, the larger n must be before the normal distribution is an adequate approximation of the shape of sampling distribution of $\bar{y}$
- Why?

Central Limit Theorem – theoretical formulation

Let $\{X_1, X_2, \ldots, X_k, \ldots\}$ be a sequence of independent observations from one specific random process. Let and $E(X) = \mu$ and $SD(X) = \sigma$ and both be finite ($0 < \sigma < \infty$; $|\mu| < \infty$). If $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, sample-avg,

Then $\overline{X}$ has a distribution which approaches $N(\mu, \sigma^2/n)$, as $n \to \infty$.

Linear Combination

Given a collection of n random variables $X_1, \ldots, X_n$ and n numerical constants $a_1, \ldots, a_n$, the rv

$$Y = a_1 X_1 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a *linear combination* of the X_i's.

Expected Value of a Linear Combination

Let $X_1, \ldots, X_n$ have mean values $\mu_1, \mu_2, \ldots, \mu_n$ and variances of $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively

Whether or not the X_i's are independent,

$$E(a_1 X_1 + \ldots + a_n X_n) = a_1 E(X_1) + \ldots + a_n E(X_n)$$
$$= a_1 \mu_1 + \ldots + a_n \mu_n$$

Variance of a Linear Combination

If $X_1, \ldots, X_n$ are independent,

$$V(a_1 X_1 + \ldots + a_n X_n) = a_1^2 V(X_1) + \ldots + a_n^2 V(X_n)$$
$$= a_1^2 \sigma_1^2 + \ldots + a_n^2 \sigma_n^2$$

and

$$\sigma_{a_1 X_1 + \ldots + a_n X_n} = \sqrt{a_1^2 \sigma_1^2 + \ldots + a_n^2 \sigma_n^2}$$

Variance of a Linear Combination

For any $X_1, \ldots, X_n$, (dependent or independent!!!)

$$V(a_1 X_1 + \ldots + a_n X_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j)$$

If $X_i \sim D_i(\mu_i, \sigma_i), X_j \sim D_j(\mu_j, \sigma_j)$ and $f_{i,j}(x_i, x_j)$ is the joint density function of (X_i, X_j), then:

$$Cov(X_i, X_j) = E\big((X_i - \mu_i) \times (X_j - \mu_j)\big) =$$

$$\int (x_i - \mu_i)(x_j - \mu_j) f_{i,j}(x_i, x_j) dx_i dx_j = E(X_i \times X_j) - \mu_i \mu_j.$$

$$Corr(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}$$

A special case – Difference Between Two Random Variables

If $X_1 \sim D_1(\mu_1,\sigma_1), X_2 \sim D_2(\mu_2,\sigma_2)$ and $f_{1,2}(x_1,x_2)$ is the joint density function of (X_1, X_2), then :

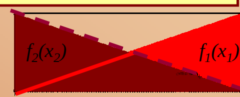
$$Cov(X_1,X_2) = E\big((X_1-\mu_1)\times(X_2-\mu_2)\big) =$$

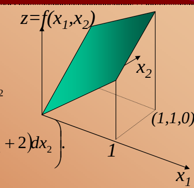
$$\int (x_1-\mu_1)(x_2-\mu_2)\ f_{1,2}(x_1,x_2)dx_1dx_2 = E(X_1 \times X_2) - \mu_1\mu_2.$$

$$Corr(X_1,X_2) = \frac{Cov(X_1,X_2)}{\sigma_1\sigma_2}$$

Example: Variance of Linear Combinations

If $X_1 \sim T_1(0,1),\ f_1(x_1) = 2x_1, x_1 \in (0,1)$
$X_2 \sim T_2(0,1),\ f_2(x_2) = -2(x_2 - 1), x_2 \in (0,1)$ and
$f_{1,2}(x_1,x_2) = x_1 + x_2, (x_1,x_2) \in (0,1)\times(0,1)$
is the joint density function of (X_1, X_2).





$$Cov(X_1,X_2) = E\big((X_1-\mu_1)\times(X_2-\mu_2)\big) =$$

$$\int_0^1 (x_1-\mu_1)(x_2-\mu_2)\ f_{1,2}(x_1,x_2)dx_1dx_2 = E(X_1 \times X_2) - \mu_1\mu_2$$

$$\int_0^1 (x_1 * x_2) * (x_1+x_2)dx_1dx_2 - \left(\int_0^1 x_1 \times 2x_1 dx_1\right)\left(\int_0^1 x_2 \times (-2x_2+2)dx_2\right).$$

$$Cov(X_1,X_2) = \frac{1}{3} - \frac{2}{3}\times\frac{1}{3} = \frac{1}{9}$$

$$Corr(X_1,X_2) = \frac{Cov(X_1,X_2)}{\sigma_1\sigma_2} = \frac{1/9}{1/2 \times 1/2} = \frac{4}{9}$$

http://www.netcomuk.co.uk/~jenolive/vect8.html

Central Limit Theorem (CLT)

Example: Applets

http://socr.stat.ucla.edu/Applets.dir/SamplingDistributionApplet.html

Application to Data

Example: LA freeway commuters (mean/SD systolic pressure):

$\mu = 130$
$\sigma = 20$

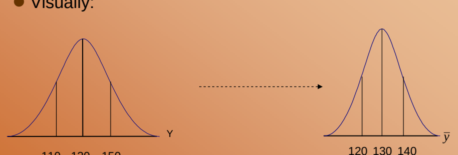
Suppose we randomly sample 4 drivers.

Find $\mu_{\bar{y}}$ $\qquad \mu_{\bar{y}} = \mu = 130$

Find $\sigma_{\bar{y}}$ $\qquad \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{4}} = 10$

Application to Data

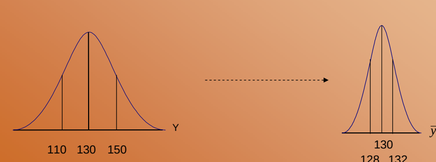
- Visually:



Application to Data

Example: LA freeway commuters (cont')

Suppose we randomly select 100 drivers



As n gets larger the variability in the sampling distribution gets smaller.

Application to Data

Example: LA freeway commuters (cont')

Suppose we want to find the probability that the mean of the 100 randomly selected drivers is more than 135 mmHg

- First step: Rewrite with notation!
  $\bar{y}$ ~ N(130,2)
- Second step: Identify what we are trying to solve!

$$P(\bar{y} > 135)$$

Application to Data

- Third step: Standardize

$$P(\bar{y} > 135) = P\left(\frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} > \frac{135 - 130}{2}\right) = P(Z > 2.5)$$

- Fourth Step: Use the standard normal table to solve
  1 – 0.9938 = 0.0062

  If we were to choose many random samples of size 100 from the population about 0.6% would have a mean SBP more than 135 mmHg.

Application to Data

Example: LA freeway commuters (cont')

n	$P(125 < \bar{Y} < 135)$	$\sigma_{\bar{y}}$
4	$P(-0.5 < Z < 0.5) = 0.3830$	$20/\sqrt{4} = 10$
10	$P(-0.79 < Z < 0.79) = 0.5704$	$20/\sqrt{10} = 6.32$
20	$P(-1.12 < Z < 1.12) = 0.7372$	$20/\sqrt{20} = 4.47$
50	$P(-1.77 < Z < 1.77) = 0.9232$	$20/\sqrt{50} = 2.83$

The mean of a larger sample is not necessarily closer to μ , than the mean of a smaller sample, but it has a greater probability of being closer to μ.

Therefore, a larger sample provides more information about the population mean

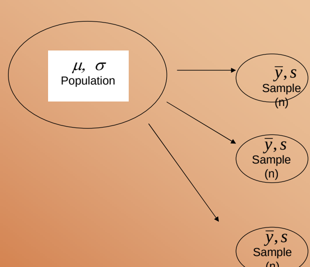
Notation

- Notation:

	mean	standard deviation
Population	μ	σ
Sample	$\bar{y}$	s
Sampling Distribution of $\bar{y}$	$\mu_{\bar{y}}$	$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

Other Aspects of Sampling Variability

- Sampling variability in the shape
- Sampling variability in the sample standard deviation



Overall: If n is large, s⟶σ, the shape of each sample will be close to the shape of the population, and the shape of the sampling distribution of $\bar{y}$ will approach a normal distribution.

Statistical Estimation

- This will be our first look at statistical inference
- Statistical estimation is a form of statistical inference in which we use the data to:
  - determine an estimate of some feature of the population
  - assess the precision of the estimate

Statistical Estimation

Example: A random sample of 45 residents in LA was selected and IQ was determined for each one. Suppose the sample average was 110 and the sample standard deviation was 10.

What do we know from this information?

$\bar{y}$ = 110

S = 10

Statistical Estimation

- The population IQ of LA residents could be described by μ and σ

110 is an estimate of μ

10 is an estimate of σ

- We know there will be some sampling error affecting our estimates
  - Not necessarily in the measurement of IQ, but because only 45 residents were sampled

Statistical Estimation

- QUESTION: How good is $\bar{y}$ as an estimate of μ?
- To answer this we need to assess the reliability of our estimate $\bar{y}$
- We will focus on the behavior of $\bar{y}$ in repeated sampling
  - Our good friend, the sampling distribution of $\bar{y}$

The Standard Error of the Mean

- We know the discrepancy between μ and $\bar{y}$ from sampling error can be described by the sampling distribution of $\bar{y}$, which uses $\sigma_{\bar{y}}$ to measure the variability
  - Recall: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
- Is there a problem with obtaining $\sigma_{\bar{y}}$ from our data?
- What seems like a good estimate for $\sigma_{\bar{y}}$?

$\frac{s}{\sqrt{n}}$ is an estimate for $\frac{\sigma}{\sqrt{n}}$

Called the *standard error of the mean*

The Standard Error of the Mean

- Notation for the standard error of the mean

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

  - Sometimes referred to as the standard error (SE)
- Round to two significant digits

The Standard Error of the Mean

Example: LA IQ (cont')

$$SE_{\bar{y}} = \frac{10}{\sqrt{45}} = 1.49$$

- What does this mean?
  - Because the standard error is an estimate of $\sigma_{\bar{y}}$, it is a measure of reliability of $\bar{y}$ as an estimate of μ.
  - We expect $\bar{y}$ to be within one SE of μ most of the time

The Standard Error of the Mean

- If SE is small we have a more precise estimate
- The formula for SE uses ***s*** (a measure of variability) and ***n*** (the sample size)
  - Both affect reliability.

Example: LA IQ (cont')

s describes variability from one person in the sample to the next SE describes variability associated with the mean (our measure of precision for the estimate)

The Standard Error of the Mean

	n	$\bar{y}$	SE	s
Male	5	117	6.40	14.3
Female	40	109	3.16	20.0

As $n \longrightarrow \infty$,

$\bar{y} \longrightarrow \mu$

$s \longrightarrow \sigma$

$SE \longrightarrow 0$

- Example: LA IQ (cont')

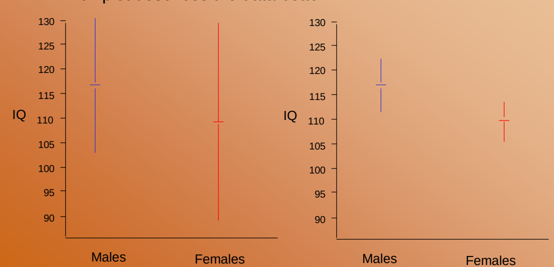
Suppose the results of the

45 LA residents were analyzed by gender.

Females have greater variability, but a much smaller SE because their sample size is larger. Therefore the females will have a more reliable estimate of μ.

The Standard Error of the Mean

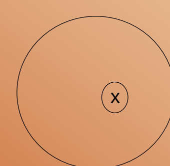
- Which plot represents the SD and the SE?
- Which plot describes the data better?



Confidence Interval for μ

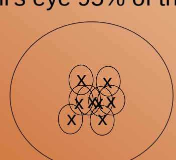
<u>Example</u>: (*Analogy from Cartoon Guide to Statistics*) Consider an archer shooting at a target. Suppose she hits the bulls eye (a 10 cm radius) 95% of the time. In other words, she misses the bulls eye one out of 20 arrows. Sitting behind the target is another person who can't see the bull's eye. The archer shoots a single arrow and it lands:

The person behind the target circles the arrow with a 10 cm radius circle, reasoning that with the archers 95% hit rate, the <u>true center</u> of bull's eye should be within <u>part</u> of that circle.



Confidence Interval for μ

As she shoots more and more arrows, the person draws more and more circles and finally reasons that these circles will include the <u>true center</u> of the bull's eye 95% of the time.



Confidence Interval for μ

- Basic idea of a confidence interval:
  - μ is the true center of the bull's eye, but we don't actually know where it is
  - We do know $\bar{y}$, which is where the arrow came through
  - We can use $\bar{y}$ and SE from the data to construct an interval that we hope will include μ

Confidence Interval for μ

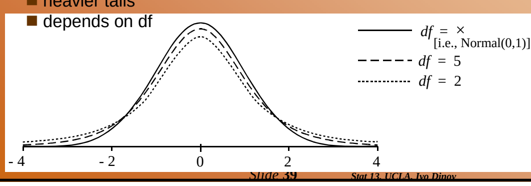
- Let's build this interval
  - From the standard normal distribution we know:
    P(-1.96 < Z < 1.96) = 0.95
  - How can we rearrange this interval so that μ is in the middle?
- Proof
- Formula $\bar{y} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$
  - will contain μ for 95% of all samples
- Any problems with using this formula with our data?
  - We can use s to estimate σ, but this changes things a little bit

The T Distribution

- If the data came from a normal population and we replace σ with s, we only need to change the 1.96 with a suitable quantity $t_{0.025}$ from the T distribution
  - Student aka William Gosset (early 1900's)
- The T distribution is a continuous distribution which depends on the degrees of freedom (df = n-1, in this case) because of the replacement we made with s: Cauchy (df=1) → T_{df} → N(0,1), df= ∞

The T Distribution

- As n approaches ∞, the t distribution approaches a normal distribution
- Similarities to the normal distribution include:
  - symmetric
  - centered at 0
- Differences from the normal distribution include:
  - heavier tails
  - depends on df



The T Table

- Table 4, p. 677 or back cover of book & Online at SOCR
- http://socr.stat.ucla.edu/Applets.dir/T-table.html
- http://socr.stat.ucla.edu/htmls/SOCR_Distributions.html
- To use the table keep in mind:
  - table works in the upper half of the distribution (above 0)
  - gives you upper tailed areas
    - this means that the "t scores" will always be positive
    - what do you do if you need a lower tail area?
  - depends on df

Using The T Table for CI's

- To use the t table for confidence intervals we will be looking up a "t multiplier" for an interval with a certain level, in this example 95%, of confidence
  - notation for a "t multiplier" is $t(df)_{\alpha/2}$
  - $t_{0.025}$ (aka $t_{\alpha/2}$) is known as "two tailed 5% critical value"
    - the interval between $-t_{0.025}$ and $t_{0.025}$, the area in between totals 95%, with 5% (aka α) left in the tails
  - If we look at the table in the back of the book we'll find:
    - $t_{0.025}$ in the 0.025 column
    - two-tailed confidence level of 95% is at the bottom of the 0.025 column
  - This is half the battle, we still need to deal with df!

Using The T Table for CI's

Example: Suppose we wanted to find the "t multiplier" for a 95% confidence interval with df = 12

$t(12)_{0.025} = 2.179$

http://socr.stat.ucla.edu/Applets.dir/T-table.html

- Recall: as n → ∞ the t distribution approaches the standard normal distribution
  - also df
  - If we look at the bottom of the table when df = ∞ , the t multiplier for a 95% CI is 1.960
    - Does anything seem familiar about this?

Calculating a CI for μ

- To calculate a $100(1 - \alpha)$ CI for μ:
  - choose confidence level (for example 95%)
  - take a random sample from the population
    - must be reasonable to assume that the population is normally distributed
  - compute: $\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

  Where $100(1 - \alpha)$ is the desired confidence
  - This means that for a 95% confidence interval α is 0.05 (or 5%, because 100(1-0.05) = 0.95

Application to Data

Example: Suppose a researcher wants to examine CD4 counts for HIV(+) patients seen at his clinic. He randomly selects a sample of n = 25 HIV(+) patients and measures their CD4 levels (cells/uL). Suppose he obtains the following results:

Descriptive Statistics: CD4

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
CD4	25	0	321.4	14.8	73.8	208.0	261.5	325.0	394.0	449.0

Calculate a 95% confidence interval for μ

Application to Data

- What do we know from the background information?

$\bar{y}$ = 321.4
s = 73.8
SE = 14.8
n = 25

$$\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 321.4 \pm t(24)_{0.05/2}\left(\frac{73.8}{\sqrt{25}}\right)$$
$$= 321.4 \pm 2.064(14.8) = 321.4 \pm 30.547$$
$$= (290.85, 351.95)$$

Application to Data

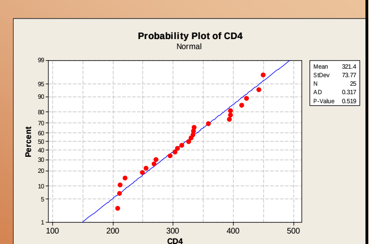
- (290.85, 351.95) – great!
- What does this mean?
  - CONCLUSION: We are highly confident at the 0.05 level (95% confidence), that the true mean CD4 level in HIV(+) patients at this clinic is between 278.58 and 342.82 cells/uL.
- Important parts of a CI conclusion:
  1. Confidence level (alpha)
  2. Parameter of interest
  3. Variable of interest
  4. Population under study
  5. Confidence interval with appropriate units

Application to Data

- Still, does this CI (290.85, 351.95) mean anything to us? Consider the following information:
  - The U.S. Government classification of AIDS has three official categories of CD4 counts –
    - asymptomatic = greater than or equal to 500 cells/uL
    - AIDS related complex (ARC) = 200-499 cells/uL
    - AIDS = less than 200 cells/uL
- Now how can we interpret our CI?

Application to Data

- Another important point to remember is that our CI was calculated assuming that the data we collected came from a population that was normally distributed!
  - N = 25 so the CLT does not protect us
  - How can we check this?

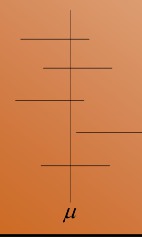


CI Interpretation

- If we were to perform a meta-experiment, and compute a 95% confidence interval about for each sample, 95% of the confidence intervals would contain μ
- We hope ours is one of the lucky ones that actually contains μ, but never actually know if it does
- We can interpret a confidence interval as a probability statement if we are careful!
  - OK: P(the next sample will give a CI that contains μ) = 0.95
    - random has happened yet
  - NOT OK: P(291 < μ < 352) = 0.95
    - not random anymore, either μ is in there or it isn't

CI Interpretation

- The confidence level is a property of the method rather than of a particular interval
- http://socr.stat.ucla.edu/htmls/SOCR_Experiments.html → CI



Other CI Levels

Example: CD4 (cont')

What if we calculate a 90% confidence interval for μ

- Without recalculating, will this interval be wider or narrower?
- NOTE: Using the same data as before, the only part that changed was the t multiplier.

95%: $t(24)_{0.025} = 2.064$

90%: $t(24)_{0.05} = 1.711$

  - As our confidence goes down the interval becomes narrower (because t gets smaller)
  - As the confidence goes up the interval becomes wider

Other CI Levels

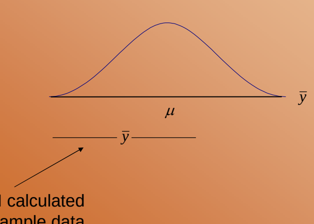
- However, we are sacrificing confidence
  - A 50% CI would be nice and small, but think about the confidence level!
- Better solution: We can also increase the sample size which will make the confidence interval narrower at the same level.
  - Why does this work?

Relationship to the Sampling Distribution of $\bar{y}$

- Recall: A CI will contain μ for 95% of samples (in repeated sampling, at 95% confidence)

$\mu_{\bar{y}} = \mu$

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



Example

Example: A biologist obtained body weights of male reindeer from a herd during the seasonal round-up. He measured the weight of a random sample of 102 reindeer in the herd, and found the sample mean and standard deviation to be 54.78 kg and 8.83 kg, respectively. Suppose these data come from a normal distribution.

Calculate a 99% confidence interval.

Example

$$\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 54.78 \pm t(101)_{0.005}\left(\frac{8.83}{\sqrt{102}}\right)$$
$$= 54.78 \pm (2.626)(0.874)$$
$$= 54.78 \pm 2.296$$
$$= (52.48, 57.08)$$

- CONCLUSION: We are highly confident, at the 0.01 level, that the true mean weight of male reindeer from the herd during this seasonal round-up is between 52.48 and 57.08 kg.

Example – 5.39 (in the textbook)

Using Binomial Counts!

- Suppose proportion of blood type **O** is 0.44. If we take a random sample of 12 subjects and make a note of their blood types what is the probability that exactly 6 subjects have type **O** blood type in the sample?
- Approach I (exact!) : P(X=6)=? Where X~B(12, 0.44) ➔

$$P(X=6) = \binom{12}{6}(0.44)^6(0.56)^6 = 0.2068 \ \ (SOCR)$$

- Approach II (Approximate): X~B(n=12, p=0.44) ➔

X (approx.) ~ N [μ = n.p = 5.28; (np(1-p))$^{1/2}$=1.7] ➔ P(X=6)~= P(Z_1<= Z <= Z_2), where Z = (X – 5.28)/1.7 and X_1=5.5, X_2=6.5

So, P(X=6)~= P(Z_1<= Z <= Z_2)=0.211

Example – 5.39 (in the textbook)

Using Proportion!

- Suppose proportion of blood type **O** is 0.44. If we take a random sample of 12 subjects and make a note of their blood types what is the probability that exactly 6 subjects have type **O** blood type in the sample?
- Approach I (exact!) : P(X=6)=? Where X~B(12, 0.44) ➔

$$P(X=6) = \binom{12}{6}(0.44)^6(0.56)^6 = 0.2068 \ \ (SOCR)$$

- Approach III (Approximate): X~B(n=12, p=0.44) ➔

p^ = X/n (approx.) ~ N [μ = p = 0.44; (p(1-p)/n)$^{1/2}$=0.1433] ➔

P(X=6) = P(p^=0.5) ~= P(p_1<= p^ <= p_2), where p_1=0.5 – 1/24 and p_2=0.5+1/24. Standardize Z = (p-0.44)/0.1433 to get:

P(X=6)~= P(p_1<= p^ <= p_2) = P(Z_1<= Z <= Z_2) = 0.211

Effect of increasing the confidence level

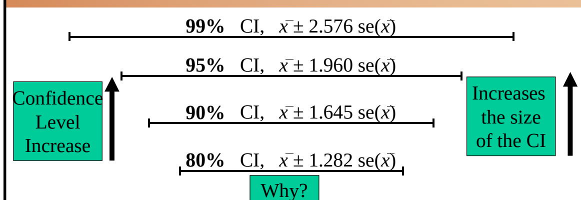
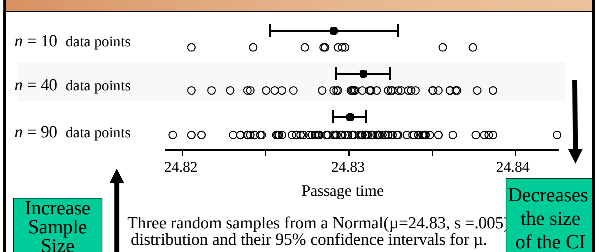


Figure 8.1.3 The greater the confidence level, the wider the interval

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

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.

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.

CI Experiment: **http://www.socr.ucla.edu/htmls/SOCR_Experiments.html**