

# Soci708 – Statistics for Sociologists

## Module 6 – Introduction to Inference<sup>1</sup>

François Nielsen

University of North Carolina  
Chapel Hill

Fall 2009

---

<sup>1</sup>Adapted in part from slides for the course Quantitative Methods in Sociology (Sociology 6Z3) taught at McMaster University by Robert Andersen (now at University of Toronto)

# Statistical Inference

- ▶ Statistical inference allows us to make statements about a population based on sample data
- ▶ It cannot tell us how correct our inferences are, but it can tell us how trustworthy they are
- ▶ Recall that statistical inference is based on probability theory:
  - ▶ The *law of large numbers* says that the larger a random sample is, the more likely the sample mean will approach the population mean
  - ▶ We use the theoretical *sampling distribution of the sample means* to judge whether our results are likely to be correct
- ▶ Finally, we can legitimately use the methods of statistical inference to be discussed here only if the data were from a *random sample or randomized experiment*

# The Sampling Distribution Revisited (1)

- ▶ The sampling distribution of a statistic (e.g.,  $\bar{x}$ ) pertains to the distribution of the statistic ( $\bar{x}$ ) for all possible samples of a given size from the population
- ▶ It is not the distribution of  $X$  in a particular sample!
- ▶ We know three facts about the sampling distribution of a sample mean ( $\bar{x}$ ):
  1. It is close to normally distributed (*central limit theorem*)
  2. The mean of all the means (i.e., the mean of  $\bar{x}$ ) equals the population mean  $\mu$
  3. The standard deviation of the mean is:

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

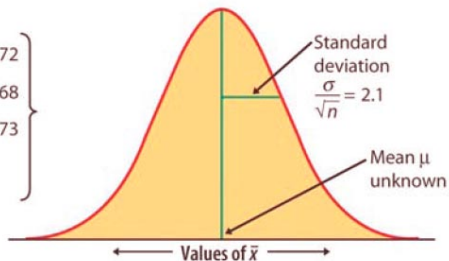
# The Sampling Distribution Revisited (2)

Situation When  $\mu$  is Unknown but  $\sigma$  Is Assumed Known (Unrealistically)



Population  
 $\mu = ?$   
 $\sigma = 60$

SRS  $n = 840$   $\bar{x} = 272$   
SRS  $n = 840$   $\bar{x} = 268$   
SRS  $n = 840$   $\bar{x} = 273$   
•  
•  
•



# Most Common Procedures of Statistical Inference

## 1. Confidence Intervals

- ▶ When we calculate a statistic, such as a mean, we have calculated a *point estimate*
- ▶ A confidence interval gives us a range around that estimate that indicates *how precisely* the estimate has been measured
- ▶ It also tells us *how confident* we can be that the population parameter lies within that range

## 2. Statistical Significance Tests

- ▶ Allow us to *test a claim* or *hypothesis* about a population parameter
- ▶ More specifically, we test whether the observed sample statistic differs from some specified value

# Statistical Significance

- ▶ An effect is considered statistically significant when it is so *large that it would rarely occur by chance alone*
  - ▶ When the probability of the result occurring by chance (called the *P-value*) is very small we say that the result is statistically significant
- ▶ We can start with a preset level of statistical significance,  $\alpha$  (alpha) that we wish to achieve
- ▶ When a P-value is less than  $\alpha$  we say that the result is statistically significant at the level  $\alpha$ 
  - ▶ It is conventional to report statistical significance cut-off levels of  $\alpha = .001$ ;  $\alpha = .01$ ; and  $\alpha = .05$
- ▶ *Statistical significance should not be confused with substantive importance!* Just because an “effect” is unlikely to be due to chance variation alone does not mean it is important

# Confidence Intervals

- ▶ Confidence intervals contain two parts:
  1. An *interval* within which the population parameter is estimated to fall

$$\text{estimate} \pm \text{margin of error}$$

2. A *confidence level* which states the probability that the method used to calculate the interval will contain the population parameter
  - ▶ If we use a *95% Confidence Interval*, we have used a method that would give the correct answer 95% of the time when using random sampling
  - ▶ In other words, 95% of samples from the sampling distribution would give a confidence interval that contains the population parameter
  - ▶ *This does not mean that the estimate is 95% correct!*

## Estimating with Confidence (1)

- ▶ Assume that we are interested in finding a 95% confidence interval for average years of education
  - ▶ This means that we want the *middle 95% of the area of the sampling distribution* for the mean of education
- ▶ Also assume that we have a SRS of 1000 people, and that the sample has mean  $\bar{x} = 12.5$ .
- ▶ Finally, assume unrealistically that we know the population standard deviation  $\sigma = 3.5$ .
- ▶ The *standard deviation of  $\bar{x}$*  is then:

$$\frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{1000}} = \frac{3.5}{31.6} = .11$$

## Estimating with Confidence (2)

- ▶ Following from the 68–95–99.7 rule, we know that for 95% of all samples  $\bar{x}$  lies within 2 standard deviations of the population mean  $\mu$ , i.e.  $\bar{x}$  lies between

$$\mu - 2 \times .11 \text{ and } \mu + 2 \times .11$$

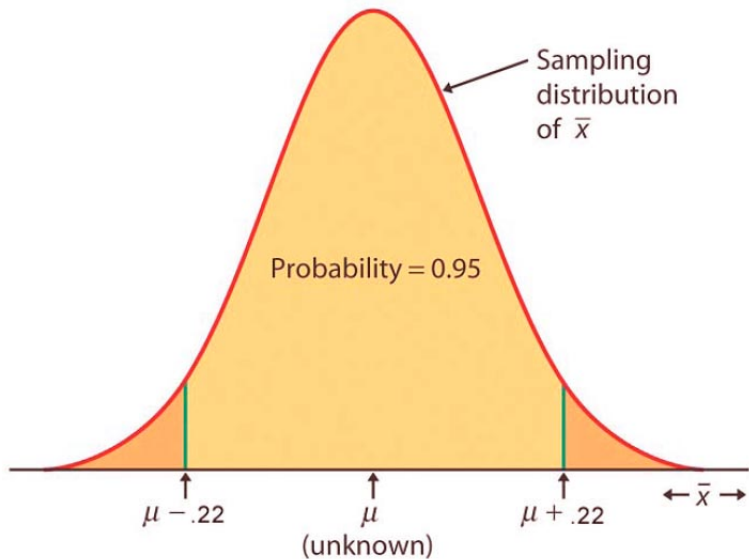
- ▶ Thus, given  $\bar{x} = 12.5$ , we know that the true population mean  $\mu$  lies somewhere between

$$\begin{aligned}\bar{x} - 2 \times .11 &= 12.5 - .22 \\ &= 12.28\end{aligned}$$

$$\begin{aligned}\text{and } \bar{x} + 2 \times .11 &= 12.5 + .22 \\ &= 12.72\end{aligned}$$

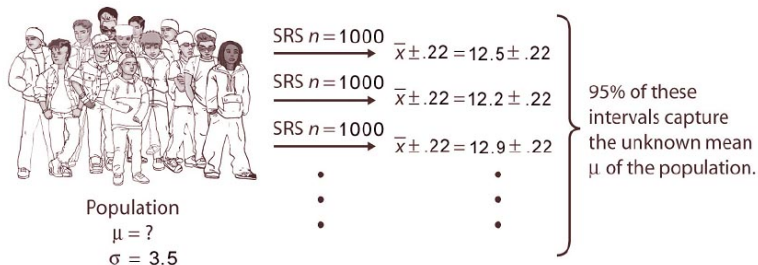
- ▶ We can say that we are 95% confident that the population mean of education is between 12.28 and 12.72

## Estimating with Confidence (3)

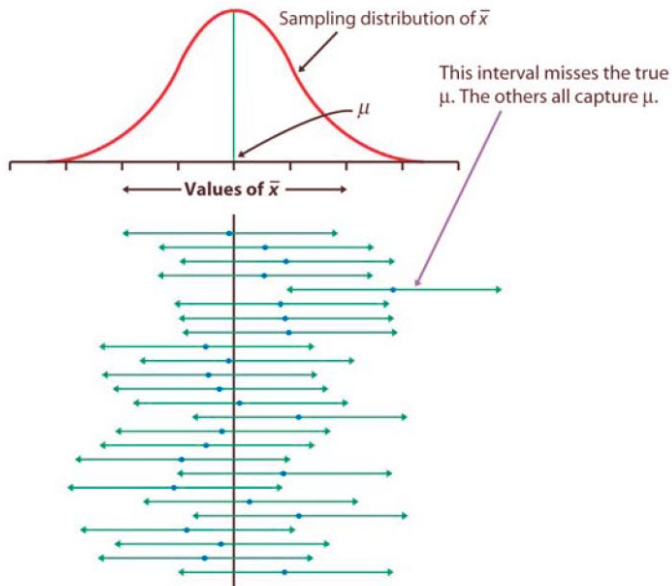


## Estimating with Confidence (4)

- ▶ Remember that although the standard deviation of  $\bar{x}$  for each possible simple random sample of size  $n$  remains constant, we would have a different mean for each sample
- ▶ In other words, our confidence interval would differ for each sample



## Estimating with Confidence (5)



## Confidence Intervals and z-scores

- ▶ Recall that when we know (unrealistically)  $\sigma$ , standardization enables us to calculate the probability of obtaining a particular result
- ▶ We can use the same procedure to standardize  $\bar{x}$  if we know  $\sigma$  (again unrealistically)
- ▶ Standardization is easily done using the following formula:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- ▶ We call this a one-sample  $z$  statistic
  - ▶ Here  $z$  tells us how far the observed sample mean ( $\bar{x}$ ) is from  $\mu$  in the units of the standard deviation of  $\bar{x}$
- ▶ Since  $\bar{x}$  is normally distributed,  $z$  is  $N(0,1)$

## Confidence Intervals and z-scores (2)

- ▶ From standardization we can solve for a confidence interval for  $\mu$ :
  - ▶ The sample mean  $\bar{x}$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ , so there is probability  $C$  that  $\bar{x}$  lies between

$$\mu - z^* \frac{\sigma}{\sqrt{n}} \text{ and } \mu + z^* \frac{\sigma}{\sqrt{n}}$$

This is exactly the same as saying that with probability  $C$  the unknown population mean  $\mu$  lies between

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \text{ and } \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

- ▶ Here  $z^*$  represents the critical value of  $z$  – i.e., the value of  $z$  that marks off the specified area  $C$  under the normal curve that is of interest

## Confidence Intervals and z-scores (3)

OPTIONAL proof of “This is exactly the same as saying ...” on previous slide

- ▶ The sample mean  $\bar{x}$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ , so there is probability  $C$  that

$$\mu - z^* \frac{\sigma}{\sqrt{n}} \leq \bar{x} \quad \text{and} \quad \bar{x} \leq \mu + z^* \frac{\sigma}{\sqrt{n}}$$

- ▶ Adding  $z^* \frac{\sigma}{\sqrt{n}}$  to both sides of the inequality on the left, one gets

$$\mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

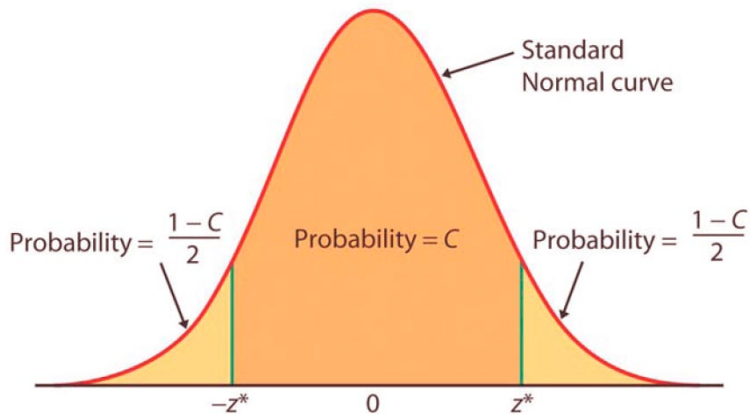
- ▶ Subtracting  $z^* \frac{\sigma}{\sqrt{n}}$  from both sides of the inequality on the right, one gets

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu$$

- ▶ Both inequalities together show that there is probability  $C$  that  $\mu$  lies between

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

The Critical Value  $z^*$  is the Value of  $z$  that Captures the Central Probability  $C$



## The Critical Value $z^*$ is the Value of $z$ that Captures the Central Probability $C$

- ▶ The critical value  $z^*$  and the central probability  $C$  are related. The most commonly used values of  $z^*$  and  $C$  are:<sup>2</sup>

$z^*$	1.645	1.960	2.576
$C$	90%	95%	99%

- ▶ E.g., the interval  $\pm 2.576$  contains 99% of the area centered around 0 under the standard normal density curve

---

<sup>2</sup>See Moore & al. 2009, Table D, bottom row

# Steps to Calculating the Confidence Interval for a Mean

## (1)

1. Calculate the point estimate for the sample mean:

$$\bar{x} = \frac{\sum x_i}{n} = 12.5$$

2. Find the critical value  $z^*$  of  $z$  that corresponds to the desired level of confidence

- ▶ For example, a 95% CI requires the middle 95% of the area
- ▶ Therefore, we need  $z^*$  for an area

$$(1 - .95)/2 = .025 \text{ and } 1 - .05/2 = .975$$

- ▶ We see from Table A or Table D (bottom row) that  $z^* = \pm 1.96$
3. Substitute the known values into the formula:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

## Steps to Calculating the Confidence Interval for a Mean (2)

- ▶ Continuing from the education example, we know the following:

$$n = 1000$$

$$\sigma = 3.5$$

$$\bar{x} = 12.5$$

$$z^* = 1.96$$

- ▶ Therefore,

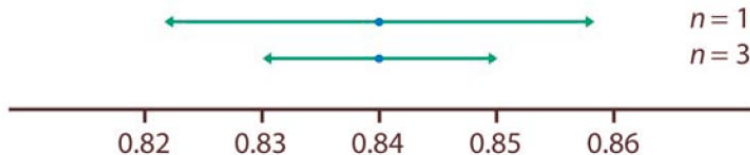
$$\begin{aligned}\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 12.5 \pm 1.96 \frac{3.5}{\sqrt{1000}} \\ &= 12.5 \pm .217 \\ &= 12.283 \text{ to } 12.717\end{aligned}$$

- ▶ We can conclude that we are 95% confident that the true education mean lies between 12.283 and 12.717 years

## Sample Size and Margin of Error

- ▶ Recall that standard deviation of the sampling distribution gets smaller as the size of the sample gets larger
  - ▶ As a result, as the size of the sample gets larger, the margin of error gets smaller
  - ▶ In other words, statistics become more precise as the sample size gets large
- ▶ Notice below that the sample size appears in the denominator for the equation for the margin of error

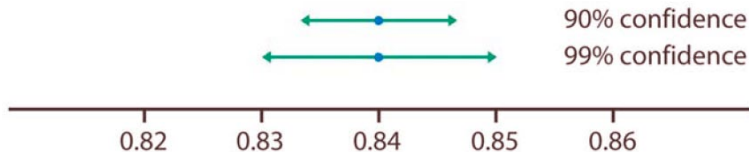
$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$



## Confidence Level and the Margin of Error

- ▶ When the confidence level gets larger, the critical value  $z^*$  gets larger
- ▶  $z^*$  is on the numerator for the formula for the margin of error
  - ▶ As a result, as the confidence level gets larger, the margin of error also gets larger
  - ▶ In other words, the more confident we are in our results, the larger the margin of error will necessarily be

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$



## Margin of Error and Choosing the Appropriate Sample Size

- ▶ Before collecting data, it is often desirable to determine the required sample size for a specific margin of error
- ▶ Simply substitute the values for  $z^*$  (confidence level),  $\sigma$ , and the desired margin of error  $m$  in the following formula:

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

- ▶ Assume that we want a margin of error of  $m = .2$  for a 95% CI ( $z^* = 1.96$ ) and  $\sigma = 3.5$ :

$$n = \left( \frac{z^* \sigma}{m} \right)^2 = \left( \frac{1.96 \times 3.5}{.2} \right)^2 = 1176.49$$

- ▶ We conclude that we need a sample size of 1177 (*always round up*)

## Cautions Regarding Confidence Intervals

- ▶ CIs are only appropriate when the data are from a *simple random sample*
  - ▶ CIs do not apply to non-probability samples.
  - ▶ The formula discussed today must be altered for more complex sampling designs such as cluster samples
- ▶ The confidence interval can be unduly influenced by *outliers*
- ▶ The margin of error covers *only random sampling errors*
  - ▶ It does not cover errors associated with undercoverage and nonresponse – i.e., *systematic bias*

# Significance Tests (1)

- ▶ *Tests of Significance allow us to test claims about a population*
  - ▶ For example, we could test whether the mean of the population was a specified value
  - ▶ We could also test for difference between two means
- ▶ Tests of significance are based on the same principles as confidence intervals
  - ▶ Again, we use the theoretical sampling distribution to assess what would happen if we used the inference method many times
  - ▶ We can specify a significance level  $\alpha$  (alpha) beforehand, or simply judge significance based on the *P-value* (the probability of getting an outcome more extreme than the one we observe)

## Significance Tests (2)

1. We start with an *alternative hypothesis* which expresses the difference or effect we *expect* or *hope* to find in the data

$$H_a : \mu > \mu_0$$

2. The null hypothesis is then the *opposite (complement)* of the alternative hypothesis we are trying to determine

$$H_0 : \mu \leq \mu_0$$

Note that the null hypothesis includes the possibility of equality with the “null” value

- ▶ The union of  $H_a$  and  $H_0$  constitutes the sample space  $S$  (here, the real line)
3. Finally, we carry out a *test of significance*
    - ▶ If we get an outcome that is *unlikely if the null is true*, we *reject the null hypothesis*
    - ▶ Small P-values are evidence against the null

## Significance Tests: An example (1)

- ▶ Recalling the education example once again, we know the following information:

$$n = 1000$$

$$\sigma = 3.5$$

$$\bar{x} = 12.5$$

$$z^* = 1.96$$

- ▶ Assume (hypothetically) that the literature suggests that the average level of education in the population of interest is about 11 years. We think it is higher, so we set out to test the claim that  $\mu > 11$
- ▶ We start by stating the alternative hypothesis which in this case is that average education level is *greater* than 11 years:

$$H_a : \mu > 11$$

## Significance Tests: An example (2)

- ▶ We now state the null hypothesis:
  - ▶ Since we are testing whether the average education is higher, our null hypothesis is:

$$H_0 : \mu \leq 11$$

- ▶ We now carry out a z-test to see what proportion of samples would give an outcome as extreme as ours (12.5) if the null hypothesis ( $\mu = 11$ ) is correct
  - ▶ In this case we are looking for the proportion of samples that would be higher than 12.5 if  $\mu = 11$ :

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{12.5 - 11}{3.5 / \sqrt{1000}} = \frac{1.5}{.1106} = 13.553$$

- ▶ We now look to Table A in Moore & al. (2009) to find the P-value (area or probability *to the right* of the z-statistic of 13.55)

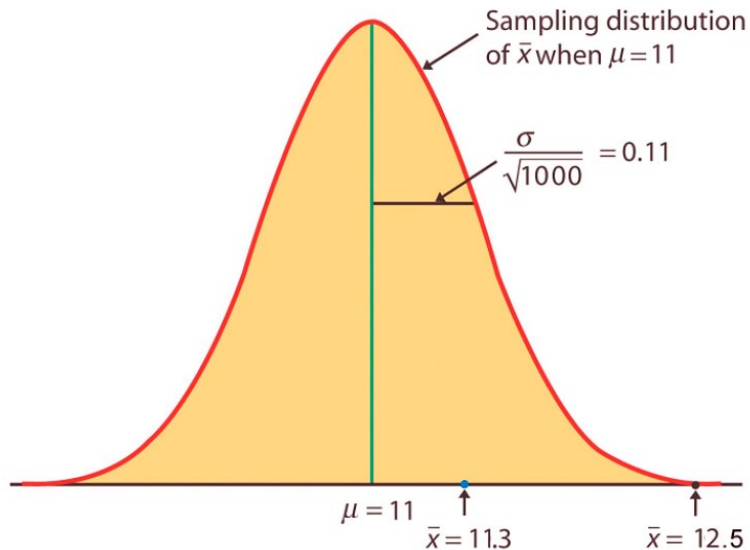
## Significance Tests: An example (3)

- ▶ We would usually report the specific P-value, but since the table does not give z-scores nearly this high, we simply say that  $P < .0002$  (since  $1 - .9998 = .0002$  is the p-value for the largest z-score in the table)<sup>3</sup>
  - ▶ A z-score this large is a highly unlikely outcome if  $\mu = 11$
- ▶ In other words, *we can reject the null hypothesis that the population mean equals 11*
- ▶ We can conclude that the average education level in the population is higher than 11 years

---

<sup>3</sup>With software we can find that the P-value  $P(Z > 13.55)$  is essentially zero

## Significance Tests: An example (4)



## One-sided Significance Tests on the Low Side

- ▶ The previous example showed us how to perform a significance test when we expect that the population mean is higher than a specified value  $\mu_0$
- ▶ We can just as easily test whether a population mean is smaller than a specified value.
- ▶ In this case we simply rewrite the alternative hypothesis:

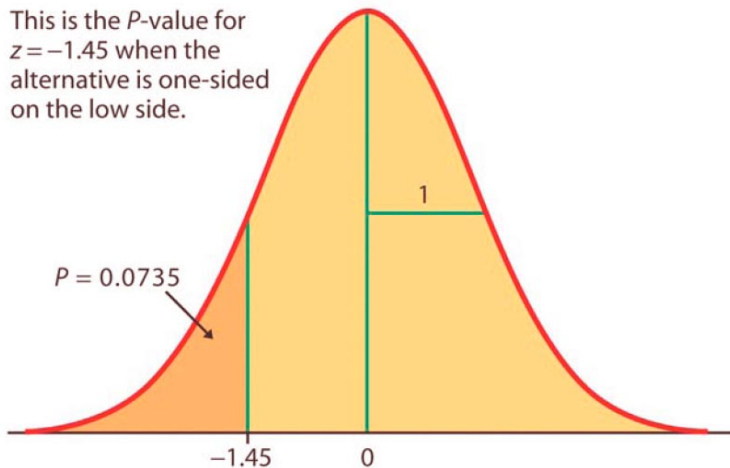
$$H_a : \mu < \mu_0$$

- ▶ We then calculate the z-statistic and find its corresponding P-value, which is now simply the area *to the left* of the z-score
- ▶ E.g., we suspect that average education is less than  $\mu_0 = 11$  years. The sample mean is  $\bar{x} = 10.83951$ . Then

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{10.83951 - 11}{3.5 / \sqrt{1000}} = -1.45$$

## One-sided Test on the Low Side

This is the  $P$ -value for  $z = -1.45$  when the alternative is one-sided on the low side.



## Two-Sided Significance Tests (1)

- ▶ The previous examples were of one-sided significance tests
  - ▶ We were interested in deviations that were different from the null hypothesis only in one direction – i.e., we tested whether the population mean  $\mu$  was higher than a stated value
- ▶ One-sided tests are useful if the theory we are testing specifies that the difference should be in one direction only
- ▶ Often, however, we don't know before looking at the data in which direction the population mean  $\mu$  might differ from a specified value
  - ▶ In such cases we need to perform a two-sided test
  - ▶ It is considered “cheating” to look at the data first – if we don't know in which direction to expect the difference, we must specify a two-sided test

## Two-Sided Significance Tests (2)

- ▶ As with the one-sided test, we start by stating the null hypothesis and the alternative hypothesis:

$$H_0 : \mu = 12.3$$

$$H_a : \mu \neq 12.3$$

- ▶ Notice here that the alternative hypothesis does not specify a direction – it only says that the population mean is not equal to  $\mu_0 = 12.3$
- ▶ As usual we calculate the z-score. Using the same sample information regarding the education example, we get:

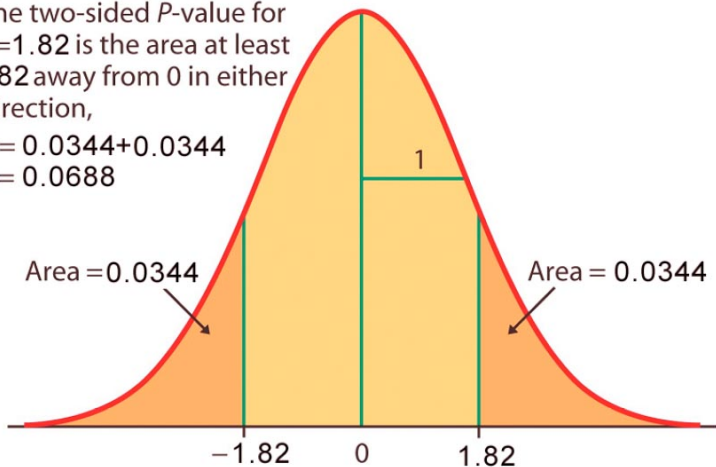
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{12.5 - 12.3}{3.5 / \sqrt{1000}} = \frac{.2}{.11} = 1.82$$

- ▶ Unlike for the one-sided test, we need to sum together the area for both above  $z = 1.82$  and below  $z = -1.82$

## Two-Sided Significance Tests (3)

The two-sided  $P$ -value for  $z=1.82$  is the area at least 1.82 away from 0 in either direction,

$$P = 0.0344 + 0.0344 \\ = 0.0688$$



# Two-Sided Significance Tests (3b)

Calculations for example in R – Effect of rounding

```
> # in R
> 12.5-12.3
[1] 0.2
> 3.5/sqrt(1000)
[1] 0.1106797
> 0.2/0.11 # rounding!
[1] 1.818182
> 2*(1-pnorm(1.82)) # rounding again!
[1] 0.068759
> # Now without rounding
> 2*(1-pnorm((12.5-12.3)/(3.5/sqrt(1000))))
[1] 0.07075981
```

- ▶ Final P-value of 0.0688 is affected by rounding of intermediate results 0.1106797 to 0.11 and 1.818182 to 1.82
  - ▶ Without rounding, the P-value would be 0.07075981
  - ▶ We would still reach the same conclusion (cannot reject  $H_0$ )

## Two-Sided Significance Tests (4)

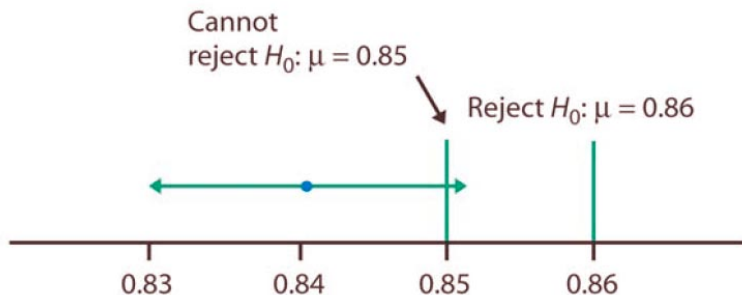
- ▶ As we saw on the previous slide, the total probability of getting a sample mean as far away from 12.3 on either side is .0688
  - ▶ From these data we would conclude that *we cannot reject the null hypothesis*. This means that we cannot say that the population mean is not 12.3
  - ▶ There is almost a 7% chance that we would get a sample mean as far away from 12.3 as we find if 12.3 is in fact the population mean
- ▶ As said earlier, it is conventional to report statistical significance cut-off levels of  $\alpha = .001$ ;  $\alpha = .01$ ; and  $\alpha = .05$ 
  - ▶ Although these conventional cut-offs are often mandated by publishing usage, it is best to report the actual P-value and let the reader decide on the statistical significance

# Confidence Intervals as Significance Tests

- ▶ A confidence interval can be used as a two-sided significance test
  - ▶ A 99% confidence interval can be used in place of a 1% significance level test
  - ▶ More generally, a two-sided test with significance level  $\alpha$  can be done directly from a confidence interval with confidence level  $C = 1 - \alpha$
- ▶ If the interval includes the specified value of interest, we cannot reject the null hypothesis
- ▶ If the interval does not contain the value, we can reject the null

## Confidence Intervals as Significance Tests

- ▶ Assume that we calculate a 99% confidence interval of  $.8404 \pm 0.0101 = .8303$  to  $.8505$
- ▶ We could use this interval to test whether  $H_0 : \mu = .86$
- ▶ Since the interval does not contain the hypothesized value, we can reject the null hypothesis at  $\alpha = .01$



## Type I and Type II Errors (1)

- ▶ When the null hypothesis  $H_0$  is wrong and we reject it, we have made the correct decision
- ▶ When the null hypothesis is right and we accept it, we have made the correct decision
- ▶ If we reject the null hypothesis  $H_0$  when it is in fact true, however, we have committed a *Type I error*
  - ▶ The probability of committing a Type I error is simply the chosen significance level  $\alpha$
- ▶ If we fail to reject the null hypothesis  $H_0$  when it should be rejected (i.e.,  $H_a$  is true) we have committed a *Type II error*
  - ▶ The probability of Type II error, denoted  $\beta$ , depends on the exact nature of  $H_a$  (e.g., the actual value of  $\mu$ )
  - ▶ The *power of a test* tells the probability that the test (for a given  $\alpha$  level) will *not* produce a Type II error
  - ▶ Power is determined simply by subtracting the probability of a Type II error from 1, i.e. Power =  $1 - \beta$

## Type I and Type II Errors (2)

**Probabilities of error vs. correct decision given the true state of the population**

Decision	True state of the population	
	$H_0$ true	$H_a$ true
Reject $H_0$	$\alpha$ Type I error	$1 - \beta$ Correct decision <sup>4</sup>
Accept $H_0$	$1 - \alpha$ Correct decision	$\beta$ Type II error
Total	1.0	1.0

<sup>4</sup> $1 - \beta$  is called the *power* of the test

## Conclusions (1)

- ▶ *Confidence intervals* give us a range (margin of error) within which the population parameter is likely to fall. They also tell us how likely our method is to work
  - ▶ We choose either the margin of error or the confidence level
- ▶ *Significance tests* allow us to test whether the population parameter is likely to be a particular value
  - ▶ We can test whether it is lower, higher or both
  - ▶ We can either choose the  $\alpha$  level ahead of time, or simply calculate the probability that our result is the same as the population parameter and make a decision based on the P-value as to whether we should accept or reject  $H_0$

## Conclusions (2)

- ▶ A confidence interval can also be used as a significance test
  - ▶ If the interval contains the value being tested by the null hypothesis, we cannot reject the null
  - ▶ If the interval does not contain the null value, we can reject the null
- ▶ Type I error occurs when we reject the null  $H_0$  even though it is true
- ▶ Type II error occurs when we accept the null even though the alternative hypothesis  $H_a$  is true
  - ▶ The *power* of a significance test tells us the ability of the test to detect an alternative hypothesis; it depends on the actual nature of  $H_a$ , e.g. the actual value of  $\mu$
- ▶ *Statistical significance* should not be confused with substantive importance

# Statistical Inference

- ▶ Next week:
  - ▶ Inference for means and proportions

# Estimating With Confidence

Practical – IPS6e 6.20 p. 370

**Apartment rental rates.** You want to rent an unfurnished one-bedroom apartment in Boston next year. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is \$1400. Assume that the standard deviation is \$220. Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community

# Tests of Significance

Practical – IPS6e 6.43 p. 385

**Computing the test statistic and  $P$ -value.** You will perform a significance test of  $H_0 : \mu = 25$  based on an SRS of  $n = 25$ .

Assume  $\sigma = 5$ .

- (a) If  $\bar{x} = 27$ , what is the test statistic  $z$ ?
- (b) What is the  $P$ -value if  $H_a : \mu > 25$ ?
- (c) What is the  $P$ -value is  $H_a : \mu \neq 25$ ?

# Tests of Significance

Practical – IPS6e 6.46 p. 388

**More on two-sided tests and confidence intervals.** A 95% confidence interval for a population mean is (57, 65).

- (a) Can you reject the null hypothesis that  $\mu = 68$  at the 5% significance level? Explain.
- (b) Can you reject the null hypothesis that  $\mu = 62$  at the 5% significance level? Explain.