# Important Points

1. how precisely we can make inferential statements about the mean of the sampled population depends on two things:
   a) how much variability (e.g., "noise") is in the data
   b) how much data was collected (i.e., # of subjects)

2. in terms of yes-or-no conclusions, there are four possible outcomes from any given experiment and two of these conclusions are errors

3. for a fixed number of subjects, one type of design makes one type of error much less often

# Statistical Conclusion Validity

- *what could go wrong (when making yes-or-no decision)?*

  the answer depends on what you concluded

- if you concluded that a difference exists (in the pop)

  then you might be making a "false-alarm" error

  which is a Type-I error

  which happens 5% of the time (since $\alpha = .05$)

- if you concluded that there isn't a difference

  then you might be making a "miss" error … Type-II

  which happens an unknown % of the time ($\beta$)

# All of the Standard Symbols

- $\mu$ (mu): population mean [unknown … usual target]
- $\bar{X}$ (X-bar): sample mean
- $\hat{\mu}$ (mu-hat): best guess for population mean = $\bar{X}$
- sd (*use full name*): standard deviation
- se (*use full name*): standard error of the mean = $sd/\sqrt{N}$
- df (*use full name*): degrees of freedom
- $\alpha$ (alpha): "significance" cut-off = .05 in psychology
- $1-\beta$ (one minus beta): "power"; aim for .80+ in psych

# Inferential Errors (from *t*-tests)

- what causes Type-I errors?
  1) bad luck
  2) one or more assumptions were not true
- what causes Type-II errors?
  1) bad luck
  2) one or more assumptions were not true
  3) noisy data and/or too small of a sample
- what are the (main) assumptions?
  1) the sampling distribution of the mean is Normal
  2) when 2+ samples, equal standard deviations

# Power and Design-type

- what could go wrong and the required assumptions are roughly the same for both types of design
- the main difference between the two types of ***t***-test (from a statistical point of view) is the size of β

  independent-samples ***t***-tests have much less power (i.e., higher β) for a given number of subjects

  1) one less degree of freedom [negligible]
  2) half as much data per condition; therefore, the $\sqrt{N}$ for calculating the se is smaller [moderate]
  3) differences between subjects adds "noise" [huge]

# Influence of Spread and Sample Size

Standard Error (of the mean) – *a measure of how far any given sample mean might be from the mean of the sampled population*

= sd / √ N

note: how more spread in sample → more error

and: larger sample → less error

therefore: you can compensate for a "noisy" DV
by collecting more data
… but this can be time-consuming (√N)

# Hypothesis Testing (yes-or-no conclusions)

| what we conclude | what is true: the pop means are the same | what is true: the pop means are different |
| --- | --- | --- |
| the pop means are the same | | |
| the pop means are different | | |

# Design Sensitivity

within-subject designs have much more power
(for a fixed number of subjects or fixed amount of data)

mostly because the (overall) differences between the subjects are "subtracted out" (and so don't add noise)

experiments (with a given N) run using a within-subject design are much less likely to make a Type-II error

expressed in reverse: in order for a between-subject design to have as much power as a within-subject design, many more subjects will have to be run