# Simple statistics II

# Statistics has 3+ components

- *Probability calculations*
  - *Descriptive statistics*
- *Data analysis*
- *Statistical inference*
  - *Inferential statistics*
- *Models ....*

# Inferential statistics, Why?

- *Our measurements have error*
  - *Random error*
  - *Measurement error*
  - *Intervening variables*
  - *Etc.*

# Inferential statistics, Why?

- *We want to make inferences beyond our sample*
- *Statistics organizes & set the "rules" by which we can draw conclusions*
- *We usually test things we think will "work"*
  - *Statistics help protect us against ourselves*

# Going beyond descriptions

- ***The main issue is variance!***
- ***The question we ask is how large or likely is the effect relative to the variance we have.***

# Sampling & probability

- *In Binomial distributions there are two possible outcomes.*
  - *What is the probability for 5 boys*
  - *What is the probability for 4 out of 5 being boys?*
- *P(r successes) = (n!/r!) * $p^r$ * $q^{n-r}$*

# Hypothesis testing #1

- *Using the binomial distribution*
- *If a family has 4 boys, are they likely to have a boy or girl next time?*
- *What about 5 or 6 boys?*

# From binomial to normal

*As $N$ increases and $p = q$, the binomial becomes close to the normal*

# Another test

- *Usually 6% of MIT students pass 15.301.*
- *At Sloan (out of 400 students) 42 have passed 15.301.*
- *Is this random? Are the Sloan students better?*

# What do we need for an answer:

- **Expected mean (   )= np**
- **Variance ($\sigma^2$) = npq**
- **Z = (xi -    ) / $\sigma$**
-
- **   = 400*6/100 = 24; $\sigma$ = 4.8**
- **Z (41.5)  = (41.5 - 24)/4.8 = 3.64**
- **Using the normal table, z = 3.64 = p 0.0001**

# Statistical tests

- T-test
- ANOVA
- Linear Regression
- Non-parametric tests

# One sample t test

*Mean diff*

$$t = \frac{- M}{\sqrt{\frac{\Sigma(xi - \quad)^2}{n-1}} \Big/ \sqrt{n}}$$

*Standard deviation*

# What do you do with “t”

- *Compare it to the “t table”*
- 
- *When there is more data, the t distribution gets closer to normal*

# Example:

| Observation | Aggressive | xi - □ | (xi - □)$^2$ |
|---|---|---|---|
| 1 | 24 | 4 | 16 |
| 2 | 22 | 2 | 4 |
| 3 | 23 | 3 | 9 |
| 4 | 18 | -2 | 4 |
| 5 | 17 | -3 | 9 |
| 6 | 16 | -4 | 16 |
| 7 | 20 | 0 | 0 |
| all | 140 | 0 | 58 |

# Example:

- *H0: average is 16*
- *H1: average ≠16*

$$\sigma = \sqrt{\frac{\sum (xi - \quad )^2}{n-1}} = 3.11$$

$$t = \frac{- M}{\sigma / \sqrt{n}} = 3.42$$

# two samples t test

## Test for independent samples

$$t = \frac{(\quad 1 - \quad 2) - (M1 - M2)}{\sqrt{\frac{n1\,\sigma 1^2 + n2\,\sigma 2^2}{n1 + n2 - 2}\left(\frac{n1 + n2}{n1 \text{ x } n2}\right)}}$$

# Example

- Who eats more lollipops males of females?
- 7 females; 5 males followed for a month
    - Females: $= 27, \sigma^2 = 29.2$
    - Males: $= 19, \sigma^2 = 24.57$
    - 
- Is there a difference?

# Calculating ...

$$t = \frac{(27 - 19) - (0)}{\sqrt{\frac{5 \text{ x } 24.57 + 7 \text{ x } 29.2}{5 + 7 - 2}\left(\frac{5 + 7}{5 \text{ x } 7}\right)}}$$

$$= 2.42$$

# two samples t test

Test for dependent samples

$$t = \frac{\text{(within diff) – (expected diff)}}{\text{sd of diff} \,/\, \sqrt{\text{n}}}$$

# Example

- Does the sun creates freckles?
- Each ss has one side of the body in the sun
-
- H0 sun side ≤ non-sun side
- H1 sun side > non-sun side

# Data

| Subject | sun | shade | diff | d - □ | (d - □)$^2$ |
|---|---|---|---|---|---|
| 1 | 6 | 8 | -2 | -3 | 9 |
| 2 | 12 | 5 | 7 | 6 | 36 |
| 3 | 3 | 2 | 1 | 0 | 0 |
| 4 | 4 | 6 | -2 | -3 | 9 |
| 5 | 7 | 0 | 7 | 6 | 36 |
| 6 | 9 | 10 | -1 | -2 | 4 |
| 7 | 4 | 4 | 0 | -1 | 1 |
| 8 | 0 | 2 | -2 | -3 | 9 |
| 9 | 4 | 3 | 1 | 0 | 0 |
| all | | | **9** | **0** | **104** |

# Calculating ...

$$\sigma = \sqrt{\frac{104}{8}} = 3.606$$

$$t = \frac{(1) - (0)}{3.606/\sqrt{9}} = 0.831$$

# Summary

- *t test as an example of inferential statistics*
- *Mean differences relative to variance*