Who really understands f ing ANOVA?

a primer on statistics for engineers

Sean Sketch
CHARM Lab Meeting

practical

Stats 101

What can we do with statistics?

1. **describe**

   summarize data about a population

   - mean & median
   - standard deviation & IQR

2. **infer**

   make educated guesses about the population from a sample

   - hypothesis testing
   - estimation
   - correlation
   - modeling

What questions might we ask with statistical inference?

Does the with haptics group perform better than the without haptics group?

What questions might we ask with statistical inference?

hypothesis testing

What questions might we ask with statistical inference?

hypothesis testing

How much faster does the subject complete the task when provided haptic feedback?

What questions might we ask with statistical inference?

hypothesis testing

estimation

What questions might we ask with statistical inference?

hypothesis testing

estimation

Is there a connection between finger size and task performance?

What questions might we ask with statistical inference?

hypothesis testing

estimation

correlation

What questions might we ask with statistical inference?

hypothesis testing

estimation

correlation

Can we create a model to fit the data, accounting for confounders, interaction, and randomness?

What questions might we ask with statistical inference?

hypothesis testing

estimation

correlation

modeling

What questions might we ask with statistical inference?

hypothesis testing

estimation

correlation

modeling

“All models are wrong, but some are useful.”

George E. P. Box

What do I mean by model?

continuous?
categorical?
binary?
counts?

fixed effects?
random effects?
confounding?
interaction?
transformations?

$$\text{response} = f(\text{predictors})$$

linear?
nonlinear?
piecewise?

special case

simple linear

$$y_i = \alpha + \beta x_i + \epsilon_i$$

$$\overrightarrow{y} = \mathbf{X}\beta + \overrightarrow{\epsilon}$$

$\min \|\overrightarrow{\epsilon}\|^2$

design matrix

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\overrightarrow{y}$$

special case

general linear

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

$$\overrightarrow{y} = \mathbf{X}\beta + \overrightarrow{\epsilon}$$

$\min \|\overrightarrow{\epsilon}\|^2$

generalized linear

$$\mathbf{E}(\mathrm{Y}) = \mu = g^{-1}(\mathbf{X}\beta)$$

$$g(\mu) = \mathbf{X}\beta$$

link function

generally fit via maximum likelihood

$\mathrm{Y} \sim \mathcal{N} \rightarrow y_i \in (-\infty, \infty)$	$g(\mu) = \mu$
$\mathrm{Y} \sim \mathrm{IG} \rightarrow y_i \in (0, \infty)$	$g(\mu) = \mu^{-2}$
$\mathrm{Y} \sim \mathrm{Poisson} \rightarrow y_i \in \mathbb{Z}^+$	$g(\lambda) = \ln(\lambda)$
$\mathrm{Y} \sim \mathrm{Bernoulli} \rightarrow y_i \in \{0, 1\}$	$g(p) = \ln(\frac{p}{1-p})$

What about random effects?

generalized linear mixed-effects models

fixed random

subject vs. subject
experts vs. novices
haptics vs. no haptics

$$g(\mu) = \mathbf{X}\beta + \mathbf{Z}u$$

fixed effects now take group randomness into account

fitting variance for group/ subject-wise noise

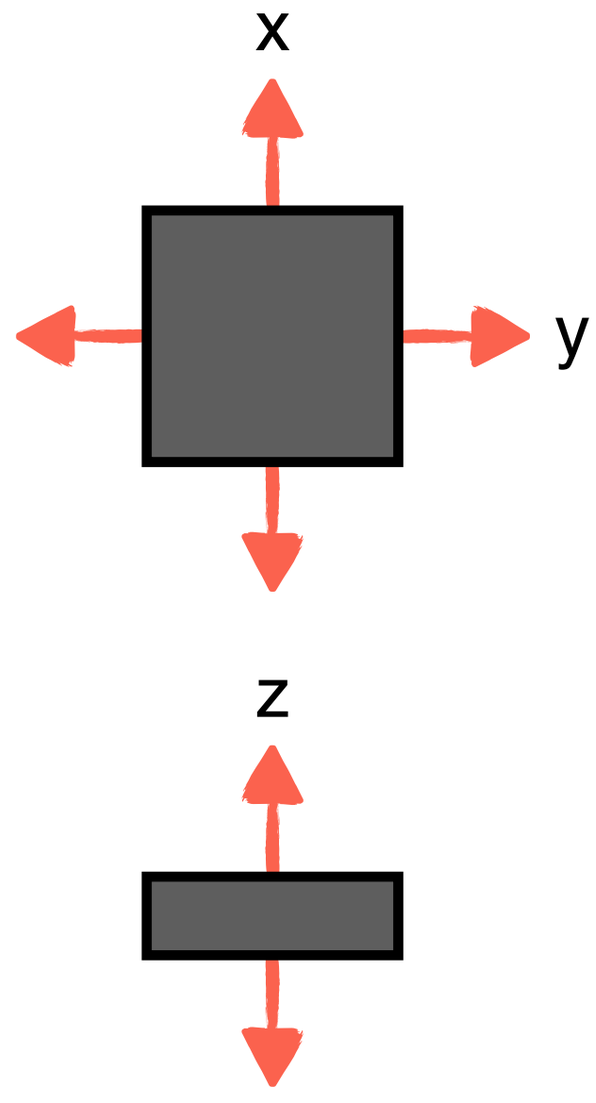
differentiating between population-average and group/ subject-specific effects — think **repeated measures**

an example

Thanks Heather!

In which directions does asymmetric vibration significantly increase the odds of correctly distinguishing that direction?

subject	stimulus #	1?	2?	3?	4?	5?	6?	# correct
1	1	1	0	0	0	0	0	5
1	2	0	1	0	0	0	0	11
1	3	0	0	1	0	0	0	7
1	4	0	0	0	1	0	0	12
1	5	0	0	0	0	1	0	5
1	6	0	0	0	0	0	1	11
2	1	1	0	0	0	0	0	12
2	2	0	1	0	0	0	0	11



an example

In which directions does asymmetric vibration significantly increase the odds of correctly distinguishing that direction?

```
# load necessary libraries
library(lme4)

# read in data
setwd("/Users/ssketch/Documents/Stanford/PhD/CHARM\ Lab/Statistics/
Examples")
vibro = read.csv("vibro.csv", header=T)

# fit logistic regression model
total = 12
model = glmer(cbind(numCorrect,total-numCorrect) ~ stimulus1 + stimulus2 +
  stimulus3 + stimulus4 + stimulus5 + stimulus6 + (1|subject),
  family = binomial, data = vibro)
summary(model)

# display fixed effects + confidence intervals, in proportion metric
exp(fixef(model))/(1+exp(fixef(model)))
exp(confint(model))/(1+exp(confint(model)))
```

an example

In which directions does asymmetric vibration significantly increase the odds of correctly distinguishing that direction?

```
> summary(model)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: cbind(numCorrect, total - numCorrect) ~ stimulus1 + stimulus2 +
    stimulus3 + stimulus4 + stimulus5 + stimulus6 + (1 | subject)
   Data: vibro

     AIC      BIC   logLik deviance df.resid
   321.8    337.1   -153.9    307.8       59

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4094 -0.9371  0.1679  1.0417  2.2416

Random effects:
 Groups  Name        Variance Std.Dev.
 subject (Intercept) 0.2277   0.4772
Number of obs: 66, groups:  subject, 11

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.28465    0.25659   5.007 5.54e-07 ***
stimulus1   -0.51822    0.28421  -1.823  0.06825 .
stimulus2   -0.08762    0.29567  -0.296  0.76697
stimulus3   -1.09196    0.27781  -3.931 8.47e-05 ***
stimulus4   -0.13020    0.29424  -0.442  0.65813
stimulus5   -0.76272    0.28035  -2.721  0.00652 **
```

an example

In which directions does asymmetric vibration significantly increase the odds of correctly distinguishing that direction?

```
> exp(fixef(model))/(1+exp(fixef(model)))
(Intercept)   stimulus1   stimulus2   stimulus3   stimulus4   stimulus5
  0.7832407   0.3732674   0.4781090   0.2512495   0.4674958   0.3180569
```

```
> exp(confint(model))/(1+exp(confint(model)))
                 2.5 %     97.5 %
.sig01       0.5662166 0.7012656
(Intercept)  0.6869090 0.8597356
stimulus1    0.2529988 0.5090431
stimulus2    0.3382458 0.6209526
stimulus3    0.1615448 0.3649667
stimulus4    0.3293194 0.6101096
stimulus5    0.2106449 0.4457801
```

statistical

FAQ

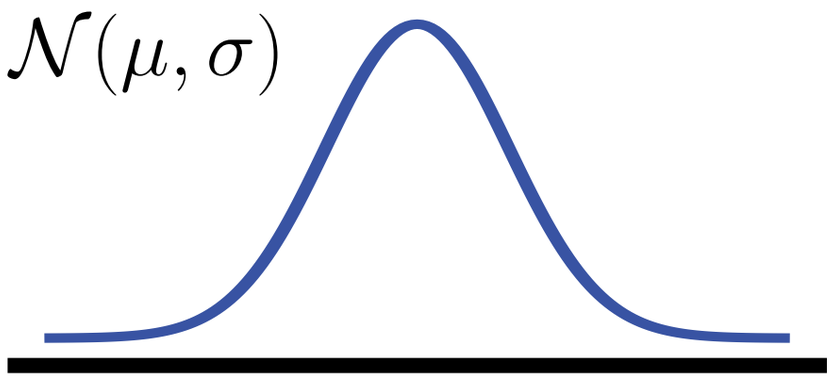
parametric **vs**. “~~non~~-parametric”

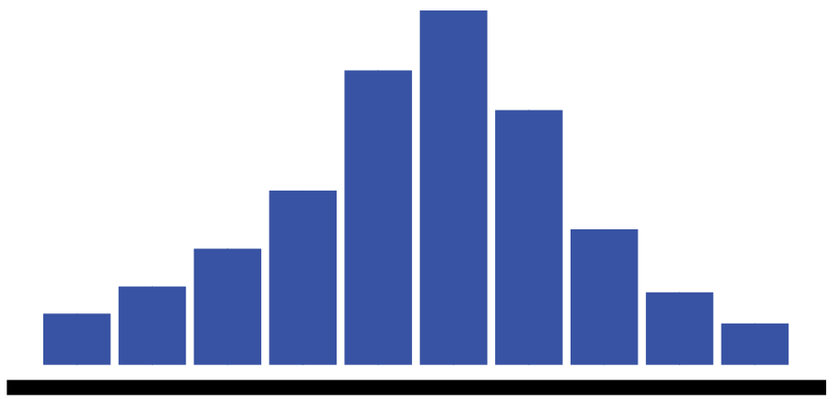
∞ ly

assumes that the variables being assessed belong to a defined probability distribution

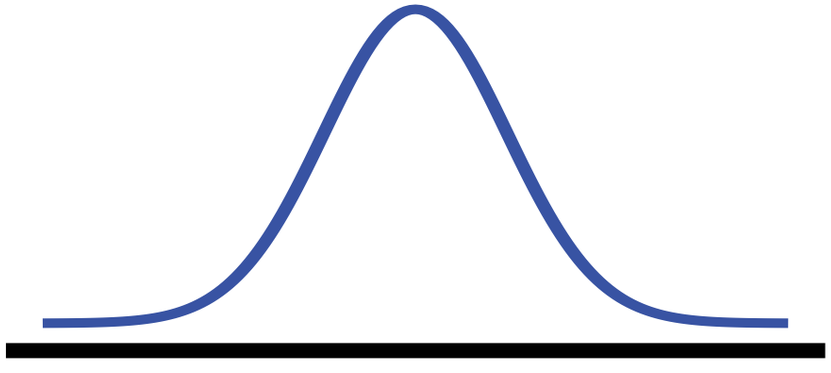
makes **NO** assumptions about the underlying probability distribution

from 2 to *n* parameters





parametric **vs**. “non-parametric”

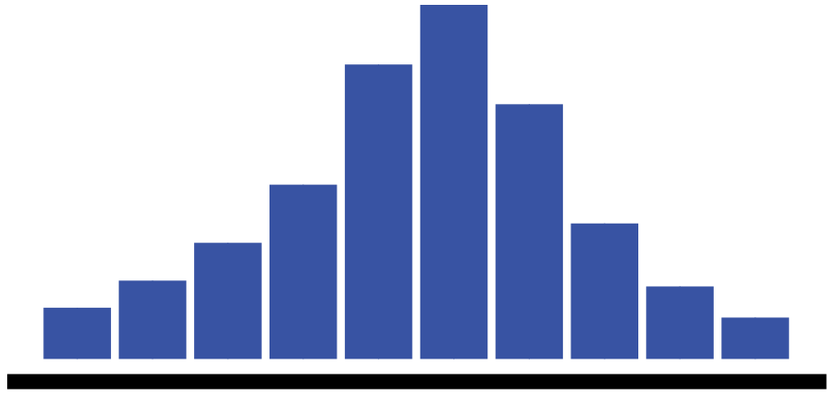


makes assumptions about underlying distributions

greater statistical power when assumptions are correct

prone to failure when assumptions are not correct

~~generally simpler & faster computations~~



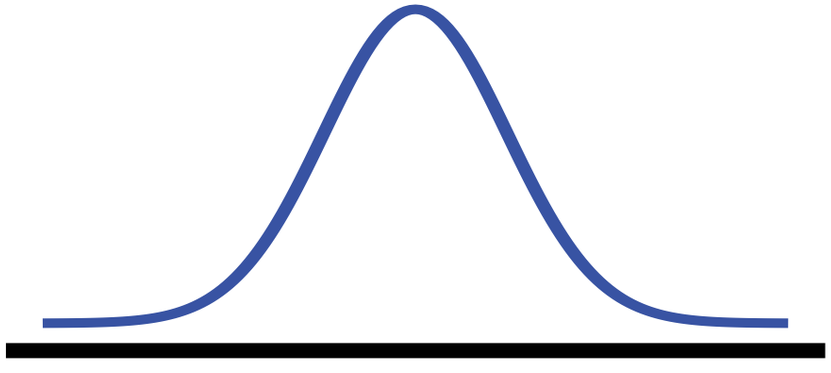
no assumptions so are always acceptable

less statistical power

~~generally more complex & slower computations~~

we have computers

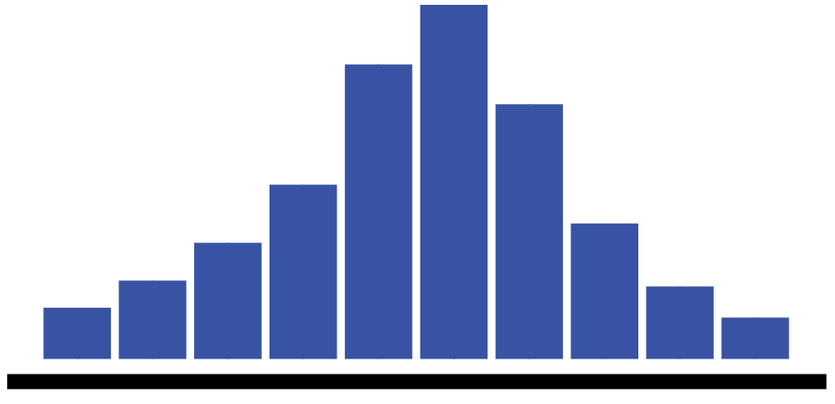
parametric **vs**. “non-parametric”



greater statistical power when assumptions are correct

prone to failure when assumptions are not correct

use when you are interested in comparing **means** and sample size is large (**n>20**)

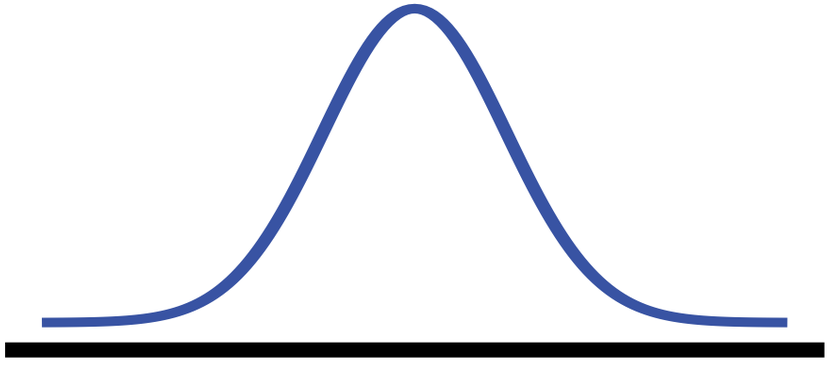


no assumptions so are always acceptable

less statistical power

use when you are interested in comparing **medians** or sample size is small (**n<20**)

parametric **vs**. “non-parametric”



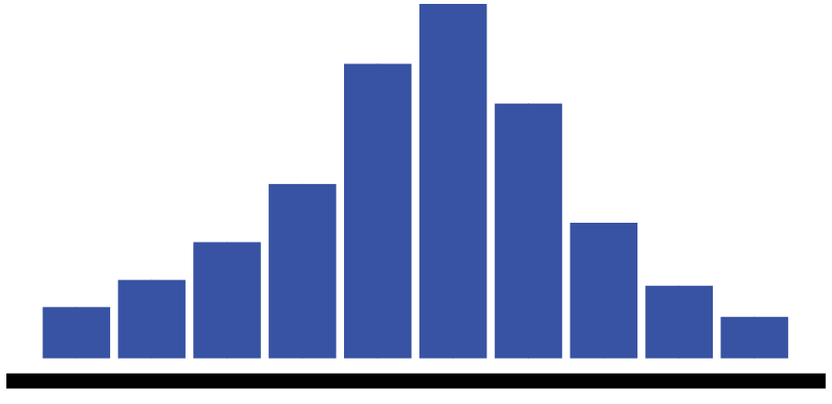
probability distribution

logistic regression

1-sample t test

2-sample t test

ANOVA
>2-sample t test



histogram

support-vector machine

Wilcoxon signed-rank test

Wilcoxon rank-sum test

Kruskal-Wallis ANOVA
>2-sample rank-sum test

bootstrapping

getting (oneself or something) into or out of a situation **using existing resources**

estimating sample statistics by drawing randomly with replacement **from existing data**

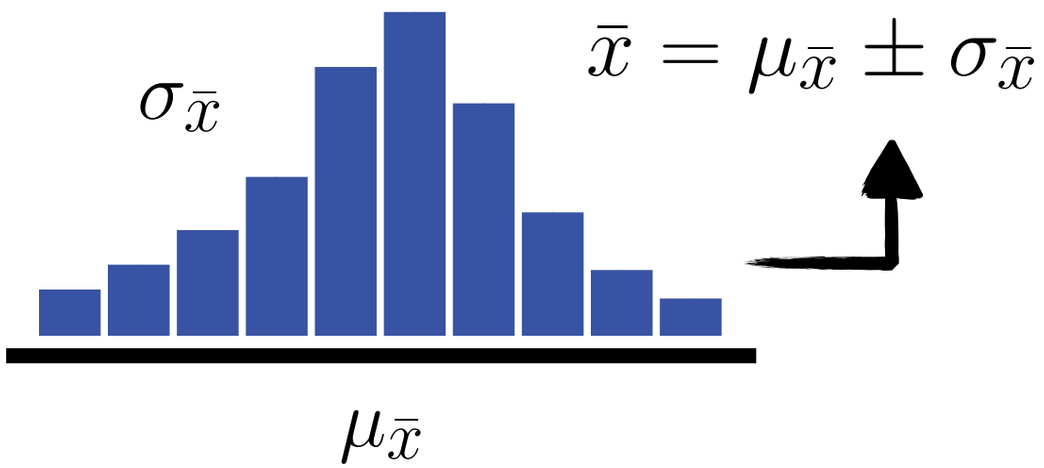
samples

x 1000

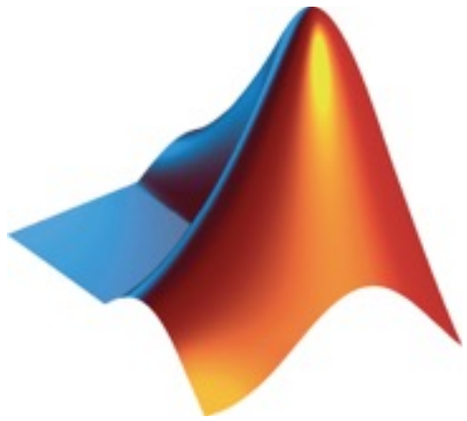
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \bar{x}$$

$n > 20$

vs.



Can I do this s[illegible] in MATLAB?



yes.

DSL designed for stats
intuitive statistical syntax (terrible otherwise)
most up-to-date & efficient packages `[lme4]`
beautiful plotting capabilities `[ggplot2]`
free & open-source

non-DSL “jack of all trades”
free & open-source

RStudio

RStudio

Go to file/function

Project: (None)

hwk4.R

Source on Save

Run

Source

```
1  # STATS 261 - Homework 4
2
3  # set working directory
4  setwd("/Users/ssketch/Documents/Stanford/Coursework/Winter\ 2017/STATS\ 262\ -\ Discrete"
5
6  # load packages
7  library(epitools)
8  library(abind)
9  library(logistf)
10 library(gtools)
11 library(pROC)
12
13 # load Kyphosis data
14 data = read.csv("kyphosis.csv")
15 attach(data)
16
```

1:1 (Top Level) R Script

Console ~/Documents/Stanford/Coursework/Winter 2017/STATS 262 - Discrete/

```
> source("roc_curve.R")
> roc.curve(glm.fit, data, "Kyphosis")

Call:
roc.formula(formula = outcome ~ prediction, data = data)

Data: prediction in 65 controls (outcome 0) < 18 cases (outcome 1).
Area under the curve: 0.9017
> neg2logL = -2*logLik(glm.fit)
>
```

Environment History

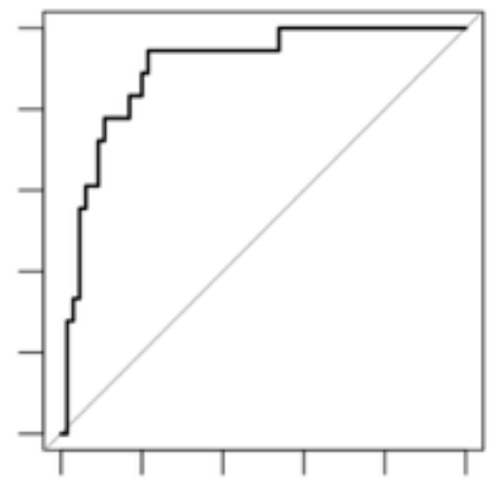
Import Dataset List

Global Environment

missing	num [1:83] 0 0 0 0 0 0 0 0…
neg2logL	Class 'logLik' : 54.62 (…
numMissing	0
results	List of 2
Functions	
logit.plot	function (data, predic…
roc.curve	function (glm.fit, dat…

Files Plots Packages Help Viewer

Zoom Export



Questions?


STANDARD MOTORCYCLES
AND SPEED SUPPLIES CO.
Octane
SUPERIOR QUALITY
NEW YORK

resources

Statistics for “Hackers” [slides & lecture]

MathWorks explanation of GLMs

parametric vs. nonparametric testing

An Introduction to the Bootstrap [Efron & Tibshirani]

R & RStudio

useful guide to R

Stanford’s biostats “core”

HRP 259: Introduction to Probability and Statistics for Epidemiology
HRP 261: Intermediate Biostatistics: Analysis of Discrete Data
HRP 262: Intermediate Biostatistics: Regression, Prediction, Survival Analysis