

Lecture 5: Statistical analyses

BIOSTAT 212, Summer 2018



Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

“mean” of a binary variable (0/1)

	id	heartatk
1	1	0
2	2	1
3	3	0
4	4	1
5	5	1
6	6	0

```
set obs 6
```

```
set seed 123
```

```
gen id = _n
```

```
gen heartatk = rbinomial(1, .5)
```

```
sum heartatk
```

sum heartatk

Variable	Obs	Mean	Std. Dev.	Min	Max
heartatk	6	.5	.5477226	0	1

Plot syntax

```
twoway                                     ///
    scatter posths acuptreatments if migraine==0    ///
,      msize(small) mcolor(orange)                ||    ///
    lfit posths acuptreatments if migraine==0        ///
,      lcolor(purple)                             ||    ///
    scatter posths acuptreatments if migraine==1    ///
,      msize(small) mcolor(blue)                   ||    ///
    lfit posths acuptreatments if migraine==1        ///
,      lcolor(green)                               ||    ///
,      title (Figure 1)
```

Plot syntax: multiple lines

twoway

```
scatter posths acuptreatments if migraine==0
, msize(small) mcolor(orange)
lfit posths acuptreatments if migraine==0
, lcolor(purple)
scatter posths acuptreatments if migraine==1
, msize(small) mcolor(blue)
lfit posths acuptreatments if migraine==1
, lcolor(green)
, title (Figure 1)
```

[illegible]

/// after each line

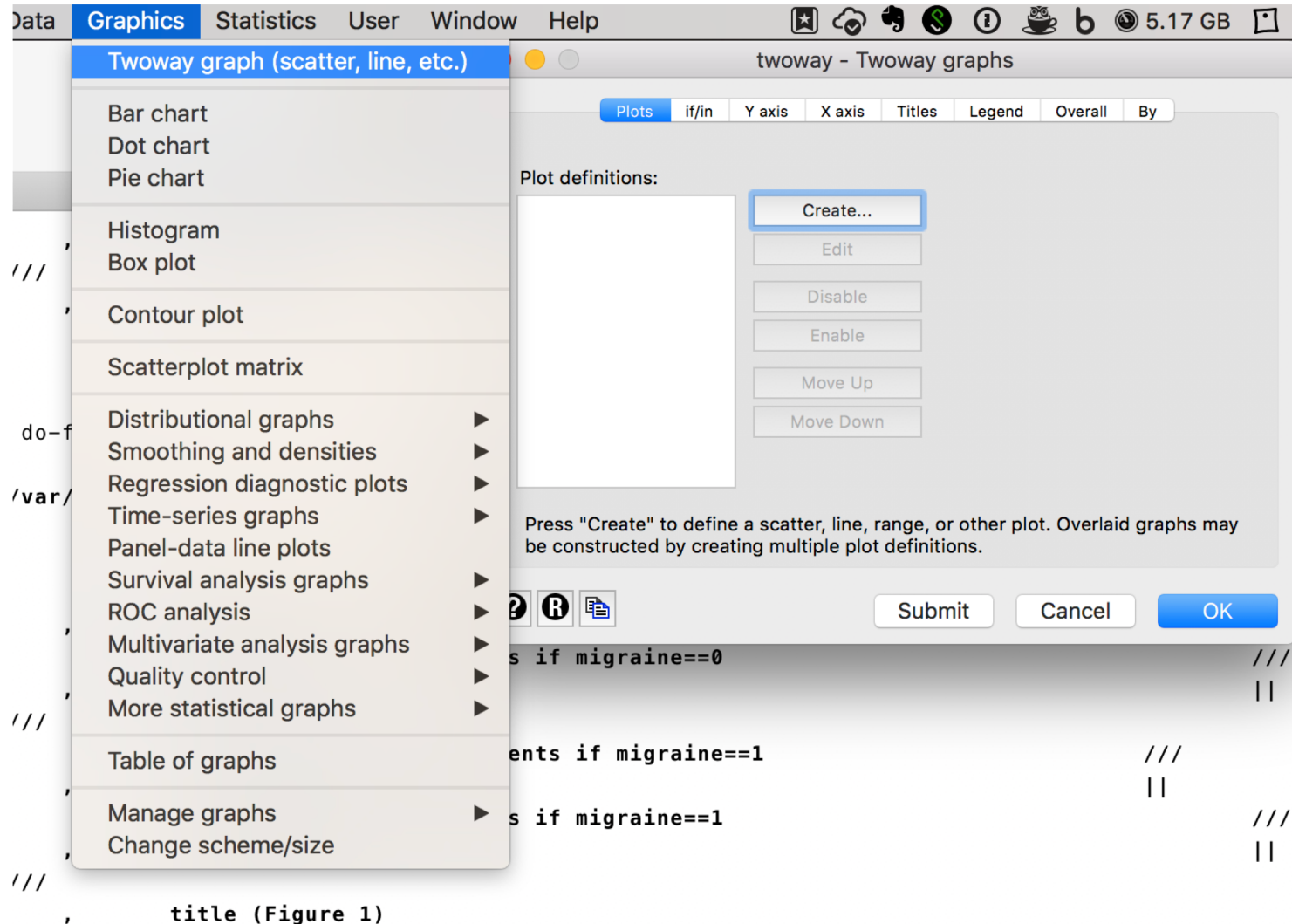
! But NOT on last line

Plot syntax: twoway

```
twoway                                     ///
    scatter posths acuptreatments if migraine==0    ///
,      msize(small) mcolor(orange)                ||    ///
    lfit posths acuptreatments if migraine==0      ///
,      lcolor(purple)                             ||    ///
    scatter posths acuptreatments if migraine==1    ///
,      msize(small) mcolor(blue)                   ||    ///
    lfit posths acuptreatments if migraine==1      ///
,      lcolor(green)                               ||    ///
,      title (Figure 1)
```

- ◇ “twoway” is optional, the plot will run without it.
- ◇ Graphics < twoway is the drop-down path to make this kind of plot

Plot syntax: twoway



Plot syntax: Layering plots

twoway

```
scatter posths acuptreatments if migraine==0
, msize(small) mcolor(orange)
lfit posths acuptreatments if migraine==0
, lcolor(purple)
scatter posths acuptreatments if migraine==1
, msize(small) mcolor(blue)
lfit posths acuptreatments if migraine==1
, lcolor(green)
, title (Figure 1)
```



///

|| after every plot

Plot syntax: Layering plots ALTERNATIVE

```
twoway                                     ///
  (  scatter posths acuptreatments if migraine==0  ///
  ,  msize(small) mcolor(orange) )           ///
  (  lfit posths acuptreatments if migraine==0      ///
  ,  lcolor(purple) )                         ///
  (  scatter posths acuptreatments if migraine==1  ///
  ,  msize(small) mcolor(blue) )             ///
  (  lfit posths acuptreatments if migraine==1      ///
  ,  lcolor(green) )                          ///
  ,  title (Figure 1)
```

() around every plot

Plot syntax: Layering plots

twoway

```
scatter posths acuptreatments if migraine==0    ///  
[ , msize(small) mcolor(orange)                ||  ///  
lfit posths acuptreatments if migraine==0       ///  
[ , lcolor(purple)                             ||  ///  
scatter posths acuptreatments if migraine==1    ///  
[ , msize(small) mcolor(blue)                  ||  ///  
lfit posths acuptreatments if migraine==1       ///  
[ , lcolor(green)                             ||  ///  
[ , title (Figure 1)
```

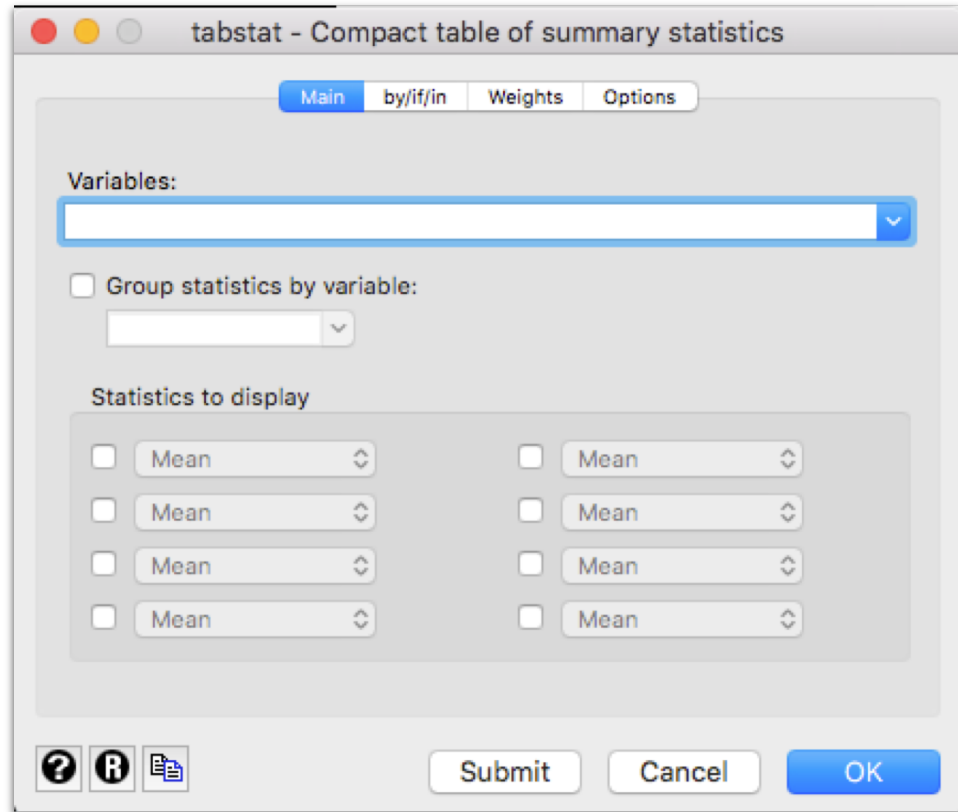
[, after EACH plot call BEFORE options

[, AND 1 comma after all plots for overall options

Outline

- ◇ **Descriptive Statistics**
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and P-values
 - T-test
 - ANOVA
 - Chi squared test
- ◇ Linear Regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

Table of summary statistics: tabstat



foreign	variable	mean	sd	min	max
Domestic	price	6072.423	3097.104	3291	15906
	weight	3317.115	695.3637	1800	4840
	mpg	19.82692	4.743297	12	34
	rep78	3.020833	.837666	1	5
Foreign	price	6384.682	2621.915	3748	12990
	weight	2315.909	433.0035	1760	3420
	mpg	24.77273	6.611187	14	41
	rep78	4.285714	.7171372	3	5

```
sysuse auto, clear
tabstat price weight mpg rep78, by(foreign)
      stat(mean sd min max) nototal long col(stat)
```

Statistics > Summaries, tables, and tests > Other tables
> Compact table of summary statistics

tabstat syntax

tabstat **varlist**, statistics(_____)

, by (varname)

-produces a summary table of statistics for all variables in **varlist**

Statistics

<u>mean</u>	-mean
<u>n</u>	-count of nonmissing observations
<u>sum</u>	-sum
<u>max</u>	-maximum
<u>min</u>	-minimum
<u>range</u>	-range = max - min
<u>sd</u>	-standard deviation
<u>variance</u>	-variance
<u>median</u>	median

*many more options-check help file

QUIZ

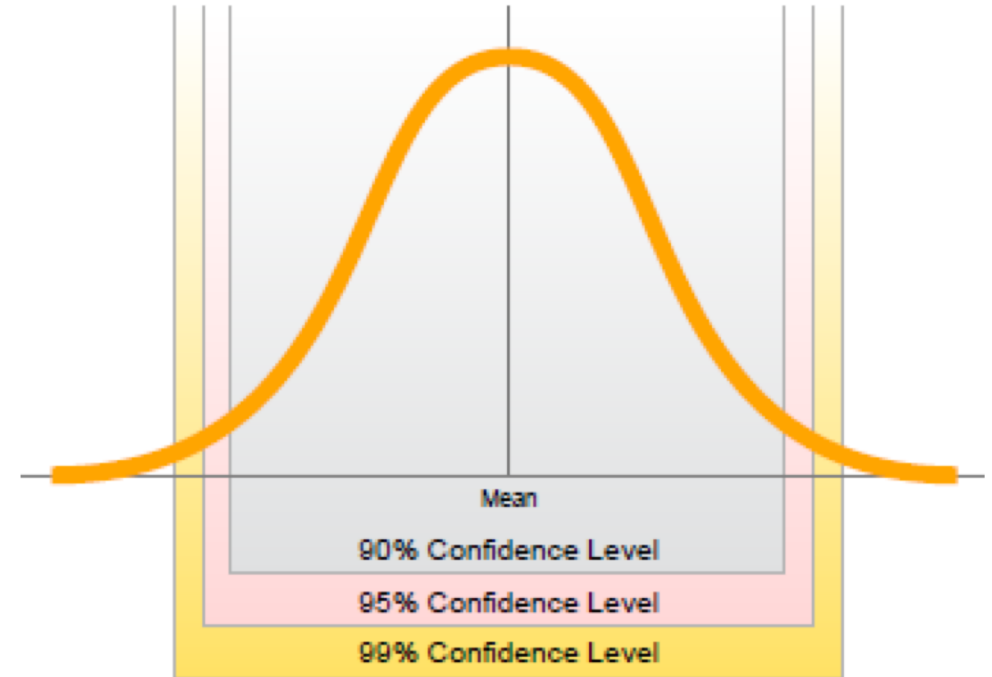
- ◇ Produce a table of [mean, standard deviation, min and max] for post-study headache score by intervention group
- ◇ What is the mean post-study headache score in the intervention group?
- ◇ What is the mean post-study headache score in the control group?

Outline

- ◇ Descriptive Statistics
- ◇ **Confidence intervals**
- ◇ Statistical hypothesis tests and P-values
 - T-test
 - ANOVA
 - Chi squared test
- ◇ Linear Regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

Confidence intervals

- Over many repetitions of the study (with different samples) 95% of estimates will be within 1.96 SEs of the true mean
 - 99% will be within 2.33 SEs



Confidence intervals: ci

ci - Confidence intervals for means, proportions, and variances

Main by/if/in Weights

Confidence intervals

☒ Means
☐ Proportions
☐ Variances
☐ Standard deviations

Variables: (leave empty for all variables)

Options

Confidence interval type

☒ Normal
☐ Poisson (counts)
☐ Poisson (rates)

Exposure variable

95 Confidence level

5 Separator line every N variables (set 0 for none)

☐ Add output for all groups combined (for use with by only)

Submit Cancel OK

```
. ci means mpg price, level(90)
```

Variable	Obs	Mean	Std. Err.	[90% Conf. Interval]	
mpg	74	21.2973	.6725511	20.17683	22.41776
price	74	6165.257	342.8719	5594.033	6736.48

```
sysuse auto  
ci means mpg price, level(90)
```

Statistics > Summaries, tables, and tests

> Summary and descriptive statistics > Confidence intervals

ci syntax

ci mean `varname`

- calculate confidence interval for a mean (continuous variable)

ci proportion `varname`

- calculate confidence interval for a proportion (dichotomous variable)

options

, level(`#`)

if

bysort `varname`:

-set confidence level (default is 95)

-subset data for CI using logic

-CI for each level of `varname`

QUIZ

- ◇ What is the 95% CIs around the mean post-study headache score in the intervention group?
- ◇ What is the 95% CIs around the mean post-study headache score in the control group?

Outline

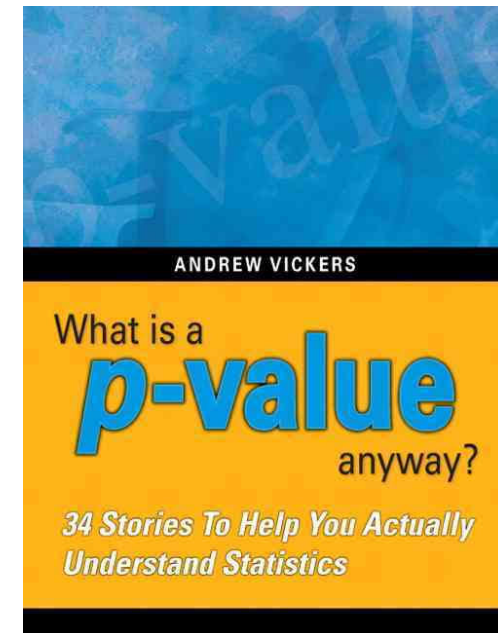
- ◇ Descriptive Statistics
- ◇ Confidence intervals
- ◇ **Statistical hypothesis tests and P-values**
 - T-test
 - ANOVA
 - Chi squared test
- ◇ Linear Regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

Hypotheses

- ◇ Null hypothesis (H_0): No association between predictor and outcome in this population
- ◇ Alternative Hypothesis (H_a): There is an association between predictor and outcome in this population
 - One-sided: states specific direction of the association (larger or smaller)
 - Two-sided: does not specify direction of the association

P-values

- ◇ The probability of obtaining a value of the test statistic equal to or one more extreme than was obtained in the study if the null hypothesis is true.



Preview: Biostat 200: Hypothesis tests

Data and comparison type	Parametric tests Stata command	Non-parametric tests Stata command
Continuous; One mean	t-test ttest varname==#	
Continuous; Two means <u>paired data</u>	Paired t-test ttest var1==var2	Sign test signtest var1=var2 Wilcoxon Signed-Rank signrank var1=var2
Continuous; Two means <u>independent data</u>	T-test ttest var1, by(var2)	Wilcoxon rank-sum test ranksum var1, by(var2)
Continuous, Two or more means <u>independent data</u>	ANOVA oneway var1 var2	Kruskal Wallis test kwallis var1, by(var2)
Dichotomous; One proportion	Proportion test prtest var1==#	
Dichotomous; two proportions	Proportion test prtest var1, by(var2)	Chi-square test tab var1 var2, chi exact
Categorical by categorical		Chi-square test tab var1 var2, chi exact

Outline

- ◇ Descriptive Statistics
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - **T-test**
 - ANOVA
 - Chi squared test
- ◇ Linear regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

t-test

1. Single sample t-test: tests the null hypothesis that the population mean is equal to a given mean
2. Paired t-test: compare 2 means when the observations are dependent (ie weight before and after intervention)
3. Independent group t-test: compare means of same variable between two groups

t-test

ttest - t tests (mean-comparison tests)

Main by/if/in

t tests

- ☒ One-sample
- ☐ Two-sample using groups
- ☐ Two-sample using variables
- ☐ Paired

One-sample mean-comparison test

Variable name: mpg Hypothesized mean: 20

95 ▼ Confidence level

Submit Cancel OK

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
mpg	74	21.2973	.6725511	5.785503	19.9569	22.63769

mean = mean(**mpg**) t = **1.9289**
Ho: mean = **20** degrees of freedom = **73**

Ha: mean < **20** Ha: mean != **20** Ha: mean > **20**
Pr(T < t) = **0.9712** Pr(|T| > |t|) = **0.0576** Pr(T > t) = **0.0288**

```
sysuse auto  
ttest mpg==20
```

Statistics > Summaries, tables, and tests > Classical tests of hypotheses > t test (mean-comparison test)

ttest syntax

- | | |
|---|----------------------------|
| <code>ttest varname==#</code> | - single sample t-test |
| <code>ttest varname1== varname2</code> | - paired t-test |
| <code>ttest varname1, by(varname2)</code> | - independent group t-test |

Options

- | | |
|-------------------------|---------------------------------------|
| <code>, level(#)</code> | – set confidence level; default is 95 |
| <code>, if</code> | – subset data using logic operators |

Practice

◇ Is mean post-study headache score different by intervention group?

- STEP 1: Visualize

```
graph bar (mean) posths, over(group)
```

- Step 2: t-test

```
ttest posths, by(group)
```

Practice

◇ Is mean baseline headache score different by sex?

- STEP 1: Visualize

```
graph bar (mean) basehs, over(sex)
```

- Step 2: t-test

```
ttest basehs, by(sex)
```


t-test: annotated output

```
ttest basehs, by(sex)
```

Two-sample t test with equal variances

Group ^a	Obs ^b	Mean ^c	Std. Err. ^d	Std. Dev. ^e	[95% Conf. Interval] ^f	
0	64	27.22656	2.317553	18.54043	22.5953	31.85782
1	337	26.37389	.8551513	15.69849	24.69176	28.05601
combined	401	26.50998	.8071529	16.16322	24.92318	28.09677
diff ^g		.8526752	2.206264		-3.48468	5.19003

diff = mean(0) - mean(1) t = 0.3865^h
Ho: diff = 0 degrees of freedom = 399ⁱ

Ha: diff < 0^k
Pr(T < t) = 0.6503

Ha: diff != 0^j
Pr(|T| > |t|) = 0.6993

Ha: diff > 0^k
Pr(T > t) = 0.3497

t-test: annotated output, cont.

- a Group** - This column gives categories of the predictor variable (sex). This variable is specified by the `by (sex)` statement.
- b Obs** - This is the number of valid (i.e., non-missing) observations in each group.
- c Mean** - This is the mean of the outcome variable (basehs) for each level of the predictor variable (sex).
- d Std Err** - This is the standard error of the mean for each level of the predictor variable (sex).
- e Std Dev** - This is the standard deviation of the outcome variable (basehs) for each of the levels of the predictor(sex). On the last line the standard deviation for the difference is given.
- f [95% Conf. Interval]** - These are the lower and upper confidence limits of the means.

t-test: annotated output, cont.

g diff - The difference in mean basehs between males and females.

h t - This is the t-statistic. It is the ratio of the mean of the difference to the standard error of the difference: $(.8526752/2.206264)$

i degrees of freedom - The degrees of freedom for the paired observations is simply the number of observations minus 2. We use one degree of freedom for estimating the mean of each group, and because there are two groups, we subtract two degrees of freedom.

j $\Pr(|T| > |t|)$ - This is the **two-tailed p-value** computed using the t distribution. It is the probability of observing a greater absolute value of t under the null hypothesis.

k $\Pr(T < t)$, $\Pr(T > t)$ - These are the one-tailed p-values for the alternative hypotheses (mean difference < 0) and (mean difference > 0), respectively.

Outline

- ◇ Descriptive Statistics
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - T-test
 - **ANOVA**
 - Chi squared test
- ◇ Linear regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

ANOVA

- ◇ The extension of the t-test to several independent groups is called analysis of variance or ANOVA (Analysis of Variance)
- ◇ The null hypothesis is:
 - H_0 : all equal means $\mu_1 = \mu_2 = \mu_3 = \dots$
- ◇ The alternative H_A is:
 - at least one of the means differs from the others

ANOVA

anova - Analysis of variance and covariance

Model Adv. model by/if/in Weights

Dependent variable:
systolic

Model:
Examples...

☐ Repeated-measures variables

Sums of squares
☒ Partial
☐ Sequential

☐ Suppress constant term
☐ Drop empty cells from design matrix

Submit Cancel OK

```
. oneway systolic drug, tabulate
```

Drug Used	Summary of Increment in Systolic B.P.				
	Mean	Std. Dev.	Freq.		
1	26.066667	11.677002	15		
2	25.533333	11.61813	15		
3	8.75	10.0193	12		
4	13.5	9.3238047	16		
Total	18.87931	12.800874	58		

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	3133.23851	3	1044.41284	9.09	0.0001
Within groups	6206.91667	54	114.942901		
Total	9340.15517	57	163.862371		

```
webuse systolic  
oneway systolic drug, tabulate
```

Statistics > Linear models and related >
ANOVA/MANOVA > Analysis of variance and covariance

ANOVA: oneway

`oneway varname1(continuous) varname2(categorical)`
– ANOVA test

Options

`, tabulate`
`, bonferroni`
`, missing`

- produces table of summary statistics
- Bonferroni multiple-comparison test
- treat missing values as a category

Practice

◇ Is mean baseline headache score different by terciles of chronicity?

- Step 1, generate chronicity tercile variable:

```
xtile chron_3 = chronicity, nq(3)
```

- Step 2: Visualize

```
graph bar (mean) basehs, over(chron_3)
```

- Now, run ANOVA:

```
oneway basehs chron_3, tabulate
```

- What is the P-value?

ANOVA: annotated output

`oneway basehs chron_3, tabulate`

3 quantiles chronicity	Summary of basehs		
	Mean ^a	Std. Dev. ^b	Freq. ^c
1	25.229167	14.606128	136
2	25.943813	16.918058	132
3	28.381579	16.848916	133
Total	26.509975	16.163224	401

Analysis of Variance					
Source	SS ^d	df ^e	MS ^f	F ^g	Prob > F ^h
Between groups	731.300961	2	365.65048	1.40	0.2472
Within groups	103768.618	398	260.72517		
Total	104499.919	400	261.249797		

Bartlett's test for equal variancesⁱ: $\chi^2(2) = 3.6110^i$ Prob> $\chi^2 = 0.164$

ANOVA: annotated output, cont.

- a Mean** - the mean of the outcome variable (basehs) for each category of the predictor variable (chronicity quantile)
- b Std Dev** - the standard deviation of the outcome variable (basehs) for each of the levels of the predictor variable (chronicity quantile)
- c Freq** - the number of observation in each category of the predictor variable (chronicity quantile)

ANOVA: annotated output, cont.

- d SS** – Sum of squares [SS between = variability around the overall mean] [SS within = weighted average of variance within each group]
- e df**- degrees of freedom [Between group: number of groups(k) – 1] [within group: total number of observations(n) – number of groups(k)]
- f MS**- mean squares is equal to the SS/df
- g F**- The F statistic is the ratio of between-group variability [MS between] to within-group variability [MS within]
- h Prob > F**- p-value for the null hypothesis test that means are the same
- i Bartlett's test for equal variance**- *ANOVA assumes that the variance is equal across groups. This test verifies that assumption. A low p-value (<0.05) indicates that the variance is not equal between groups.*

Outline

- ◇ Descriptive Statistics
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - T-test
 - ANOVA
 - **Chi squared test**
- ◇ Linear regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

Chi-squared test

- ◇ 2 categorical variables
- ◇ compares the observed frequency (O) in each cell with the expected frequency (E) under the null hypothesis of no difference
- ◇ The differences O-E are squared, divided by E, and added up over all the cells
- ◇ The sum of this is the test statistic

$$X_1^2 = \sum_{i=1}^4 \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

Chi-squared test

tabulate2 - Two-way table with measures of association

Main by/iff/in Weights Advanced

Row variable:

Column variable:

Test statistics

- ☒ Pearson's chi-squared
- ☐ Fisher's exact test
- ☐ Goodman and Kruskal's gamma
- ☐ Likelihood-ratio chi-squared
- ☐ Kendall's tau-b
- ☐ Cramer's V

Cell contents

- ☒ Pearson's chi-squared
- ☐ Within-column relative frequencies
- ☐ Within-row relative frequencies
- ☐ Likelihood-ratio chi-squared
- ☐ Relative frequencies
- ☐ Expected frequencies
- ☐ Suppress frequencies

☐ List rows in order of observed frequency

☐ List columns in order of observed frequency

☐ Treat missing values like other values

☐ Do not wrap wide tables

☐ Show cell contents key

☐ Suppress value labels

☐ Suppress enumeration log

Submit Cancel OK

Drug type (1=placebo)	1 if patient died		Total
	0	1	
1	1 5.00	19 95.00	20 100.00
2	8 57.14	6 42.86	14 100.00
3	8 57.14	6 42.86	14 100.00
Total	17 35.42	31 64.58	48 100.00

Pearson chi2(2) = 13.8678 Pr = 0.001

```
sysuse cancer  
tab drug died, row chi2
```

Statistics > Summaries, tables, and tests > Frequency tables >
Two-way table with measures of association

Chi squared test syntax

tab var1 var2, chi2

– chi squared test

Options

, row

-gives row percentages

, col

- gives column percentages

Practice

◇ Is the proportion men with migraine different from the proportion of women with migraine?

- Step 1: Visualize

```
graph bar (mean) migraine, over(sex)
```

- Step 2: Chi-squared test

```
tab sex migraine, row chi2
```


Chi squared test: annotated output

```
tab sex migraine, row chi2
```

sex	migraine		Total
	0	1	
0	5 7.81	59 92.19	64 100.00
1	19 5.64	318 94.36	337 100.00
Total	24 5.99	377 94.01	401 100.00

Pearson $\chi^2(1)^a = 0.4520^b$ Pr = 0.501^c

a (degrees of freedom) = (row-1)*(col-1)

b chi squared test statistic = sum of $[(O-E)^2/E]$

c p-value

P=0.000?

- ◇ P-value is a probability so it can't = 0
- ◇ Report as:
 - $P < 0.0005$ [gives the most information]
 - $P < 0.001$ [also ok]
 - $P < 0.01$ [also ok]
- ◇ To get the exact p-value use the test statistic:
 - `display 1 - chi2(df, chi2.statistic)`
 - `display 2*ttail(df, abs(t.statistic))`
- `df` = degrees of freedom

Outline

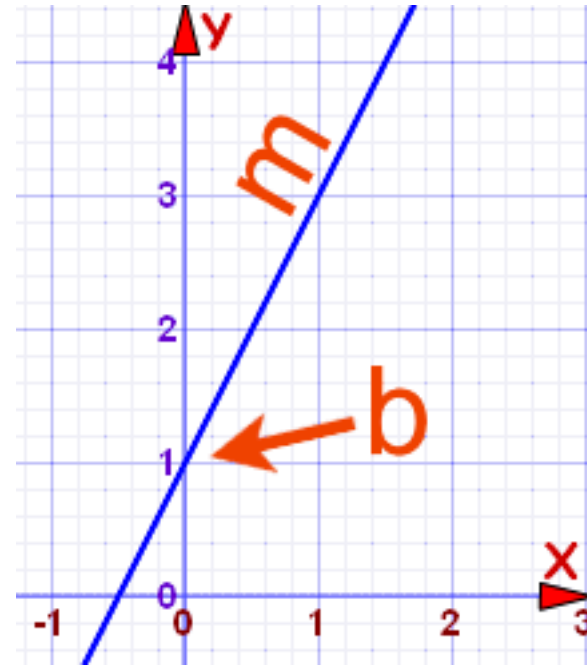
- ◇ Descriptive Statistics
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - T-test
 - ANOVA
 - Chi squared test
- ◇ **Linear regression**
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

What is a linear regression?

◇ Fit a linear equation to your data

◇ $y = mx + b$

- m =slope
- b =intercept
- y =outcome / dependent
- x =predictor / independent



Linear regression

regress - Linear regression

Model by/if/in Weights SE/Robust Reporting

Dependent variable: Independent variables:

Treatment of constant

- ☐ Suppress constant term
- ☐ Has user-supplied constant
- ☐ Total SS with constant (advanced)

Submit Cancel OK

```
. regress mpg weight
```

Source	SS	df	MS	Number of obs	=	74
Model	1591.9902	1	1591.9902	F(1, 72)	=	134.62
Residual	851.469256	72	11.8259619	Prob > F	=	0.0000
				R-squared	=	0.6515
				Adj R-squared	=	0.6467
Total	2443.45946	73	33.4720474	Root MSE	=	3.4389

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

```
sysuse auto  
regress mpg weight
```

Statistics > Linear models and related
> Linear regression

Practice

◇ Do age and chronicity (years with headache) have a linear relationship?

- Step 1: Visualize

```
scatter chronicity age || lfit chronicity age
```

- Step 2: Linear regression

```
regress chronicity age
```

Linear regression annotated output

regress chronicity age

Source ^a	SS ^b	df ^c	MS ^d	Number of obs ^e	=	401
Model	18021.6443	1	18021.6443	F(1, 399) ^f	=	124.44
Residual	57783.9268	399	144.821872	Prob > F ^f	=	0.0000
				R-squared ^g	=	0.2377
				Adj R-squared ^h	=	0.2358
Total	75805.5711	400	189.513928	Root MSE ⁱ	=	12.034

chronicity ^j	Coef. ^k	Std. Err. ^l	t ^m	P> t ^m	[95% Conf. Interval] ⁿ	
age	.6074257	.0544519	11.16	0.000	.5003772	.7144742
_cons	-6.202495	2.55145	-2.43	0.015	-11.21846	-1.18653

Linear regression annotated output

a. Source – This is the source of variance, Model, Residual, and Total. The Total variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Residual, sometimes called Error). Note that the Sums of Squares for the Model and Residual add up to the Total Variance, reflecting the fact that the Total Variance is partitioned into Model and Residual variance.

b. SS – These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual. These can be computed in many ways. Conceptually, these formulas can be expressed as:

SSTotal The total variability around the mean.

SSResidual The sum of squared errors in prediction.

SSModel The improvement in prediction by using the predicted value of Y over just using the mean of Y.

- Note that the $SSTotal = SSModel + SSResidual$.
- Note that $SSModel / SSTotal$ is equal to 0.2377, the value of R-Squared.

c. df – These are the degrees of freedom associated with the sources of variance. The total variance has N-1 degrees of freedom. In this case, there were N=401 study participants, so the DF for total is 400. The model degrees of freedom corresponds to the number of predictors + the intercept minus 1 (K-1). Including the intercept, there are 2 predictors, so the model has $2-1=1$ degrees of freedom. The residual degrees of freedom is the DF total minus the DF model, $400 - 1$ is 399.

d. MS – These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, $18021.6443 / 1 = 18021.6443$. These are computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square Residual to test the significance of the predictors in the model.

Linear regression annotated output, cont.

e. Number of obs – This is the number of observations used in the regression analysis.

f. F and Prob > F – The F-value is the Mean Square Model divided by the Mean Square Residual. The p-value is compared to your alpha level (typically 0.05) and, if smaller, you can conclude "Yes, the independent variables reliably predict the dependent variable". Note that this is an overall significance test assessing whether the group of independent variables when used together reliably predict the dependent variable, and does not address the ability of any of the particular independent variables to predict the dependent variable.

g. R-squared – R-Squared is the proportion of variance in the dependent variable (**chronicity**) which can be predicted from the independent variable (**age**). This value indicates that 23.77% of the variance in chronicity can be predicted from the variables **age**. Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

h. Adj R-squared – Adjusted R-square. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in R-square would be simply due to chance variation in that particular sample. The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population.

i. Root MSE – Root MSE is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

Linear regression annotated output, cont.

j. chronicity – This column shows the dependent variable at the top (**chronicity**) with the predictor variables below it (**age** and **_cons**). The last variable (**_cons**) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis. In other words, this is the predicted value of **chronicity** when all other variables are 0.

k. Coef. – These are the values for the regression equation for predicting the dependent variable from the independent variable. These estimates tell the amount of increase in science scores that would be predicted by a 1 unit increase in the predictor.

$$\text{Chronicity_Predicted} = -6.2 + .61 * \text{age}$$

For every year increase in age, there is a predicted .61 increase in chronicity

l. Std. Err. – These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a t-value. The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table.

m. t and P>|t| – These columns provide the t-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0.

n. [95% Conf. Interval] – This shows a 95% confidence interval for the coefficient.

Linear regression output interpretation

Coefficient for age (m) = 0.61

Intercept (b) = -6.2

Equation: $\text{chronicity} = 0.61 * \text{age} - 6.2$

For every year increase in age, there is a predicted .61 increase in chronicity.

For 50 year olds, the predicted chronicity is:

$$y = (0.6074257 * 50) - 6.202495$$

= 24.16879

`. sum chronicity if age==50`

Variable	Obs	Mean	Std. Dev.	Min	Max
chronicity	15	18.93333	14.22506	2	38

margins

- ◇ Use `margins` command to get predictions from linear model

margins

At a specific level of the predictor:

```
margins, at(age=50)
```

```
. margins, at(age=50)
```

```
Adjusted predictions      Number of obs      =      401  
Model VCE      : OLS
```

```
Expression      : Linear prediction, predict()  
at              : age              =      50
```

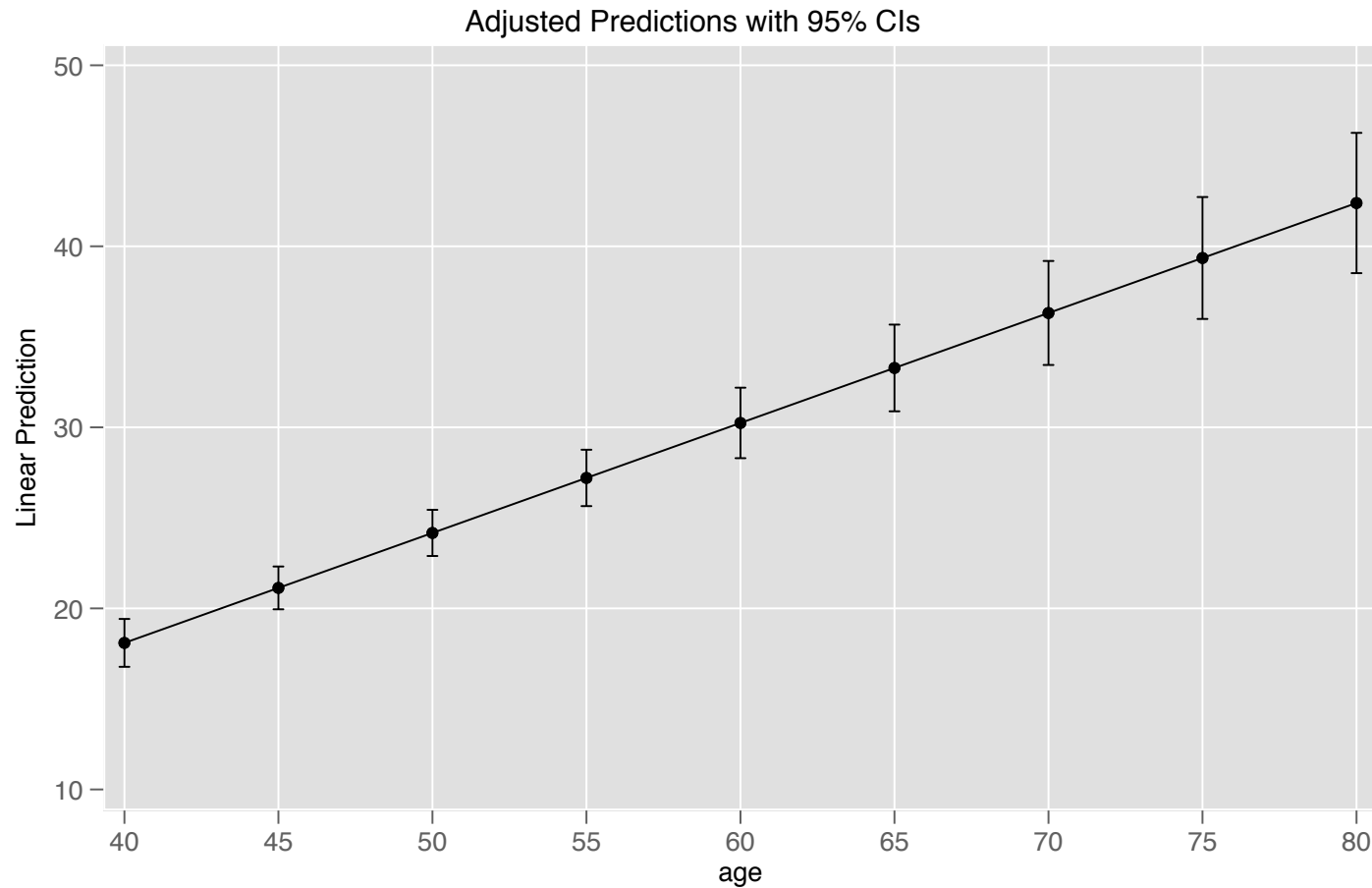
	Delta-method					[95% Conf. Interval]
	Margin	Std. Err.	t	P> t		
_cons	24.16879	.6482023	37.29	0.000	22.89447	25.44311

Or over a range of the predictor:

```
margins, at(age = (40(5)80))
```

marginsplot

- run `marginsplot` command after `margins` to visualize predictions



Outline

- ◇ **Descriptive statistics**
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - T-test
 - ANOVA
 - Chi squared test
- ◇ Linear regression
- ◇ **Immediate commands - stats**
- ◇ Power and sample size calculations

Immediate commands

- ◇ An immediate command does not require stored data to operate
 1. Calculator (display)
 - Random number generator
 2. Sample size calculator (power)
 3. Statistical/epidemiological immediate commands
 - (end in the letter i)

Immediate commands: Statistics

◇ Examples:

- `cii` confidence interval calculator

- `cii #observations #mean #standard deviation`

- `ttesti` T-test calculator

- 1 sample:

- `ttesti #obs #reported mean #sd #hypothesized mean`

- 2 sample:

- `ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2`

QUIZ

Abstract

Background: Despite high morbidity and mortality, the laboratory diagnosis of gastrointestinal infections is largely neglected in tropical African settings. This study aims to apply the Luminex multiplex PCR assay for the diagnosis of gastrointestinal pathogens in rural Ghana to evaluate its usefulness as a routine method.

Methods: A case-control study was conducted at the Agogo Presbyterian Hospital in Ghana. Stool samples were collected from children below 6 years of age with (cases) and without (controls) diarrhoea. Samples were screened for 15 different diarrhoeal pathogens by the Luminex xTAG GPP assay and associations between diarrhoea and gastrointestinal infections and fractions attributable to diarrhea (AF) were determined.

Results: The Luminex PCR assay identified organisms in 96.6 % ($n = 428$) of 443 cases and in 92.5 % ($n = 221$) of 239 selected controls. A mean of 2.5 (standard deviation [SD]: ± 1.3) and 2.3 (SD: ± 1.3) organisms per sample were detected in cases and controls respectively. An association with diarrhoea was found for rotavirus (adjusted odds ratio [aOR] = 7.2; 95 % confidence interval [CI]: 2.9–18.1), norovirus (aOR = 2.7; 95 % CI: 1.4–5.3) and *Shigella* spp. (aOR = 1.7; 95 % CI: 1.2–2.4) with respective AFs of 12.5 % (95 % CI: 9.6–15.3), 7.9 % (95 % CI: 3.8–11.7) and 16.9 % (95 % CI: 6.9–25.9).

Conclusion: The high proportion of pathogen-positive stool samples with a high number of co-infections in cases and controls suggests a substantial amount of transient or colonizing microorganisms for which treatment is not necessarily implicated. The use of sequential diagnostic algorithms with pathogen specific or quantitative PCRs might be most appropriate for diagnosing gastrointestinal infections.

Keywords: Ghana, West Africa, Multiplex Polymerase Chain Reaction, Gastrointestinal infections

1. Calculate the 95% confidence interval around the mean number of diarrhea pathogens per stool sample in cases
2. Calculate the 95% confidence interval around the mean number of diarrhea pathogens per stool sample in controls

`ci i #observations #mean #sd`

Is the mean number of pathogens different between cases and controls?

```
. ttesti 443 2.5 1.3 239 2.3 1.3
```

Two-sample t test with equal variances

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
x	443	2.5	.0617649	1.3	2.378611	2.621389
y	239	2.3	.08409	1.3	2.134344	2.465656
combined	682	2.429912	.0498772	1.30255	2.33198	2.527844
diff		.2	.1043361		-.0048597	.4048597

diff = mean(x) - mean(y)

t = 1.9169

Ho: diff = 0

degrees of freedom = 680

Ha: diff < 0

Pr(T < t) = 0.9722

Ha: diff != 0

Pr(|T| > |t|) = 0.0557

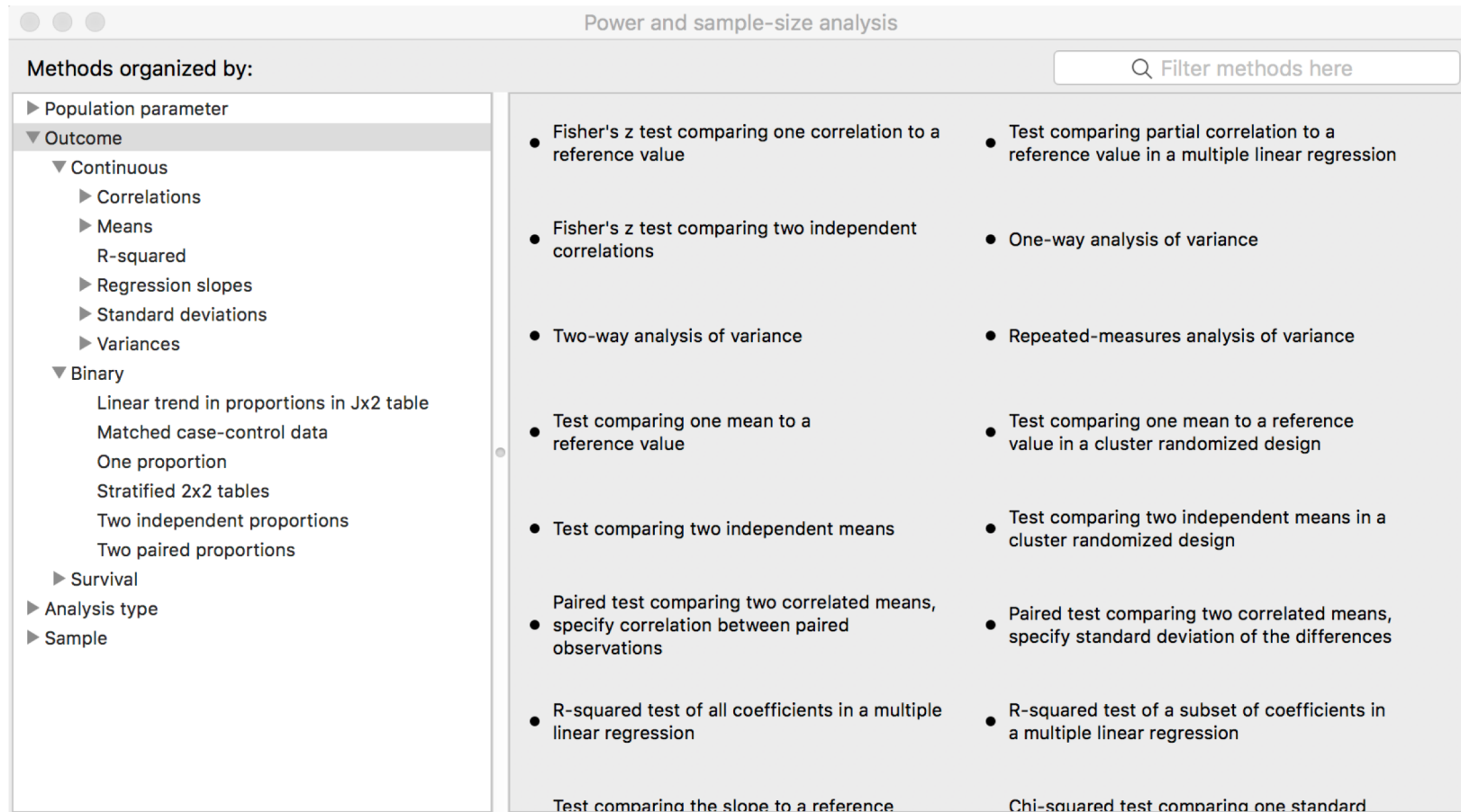
Ha: diff > 0

Pr(T > t) = 0.0278

Outline

- ◇ Descriptive statistics
- ◇ Confidence intervals
- ◇ Statistical hypothesis tests and p-values
 - T-test
 - ANOVA
 - Chi squared test
- ◇ Linear regression
- ◇ Immediate commands - stats
- ◇ Power and sample size calculations

Power and sample size calculations



Statistics < Power and sample size

What sample size would we need to evaluate the comparison of pathogens?

Abstract

Background: Despite high morbidity and mortality, the laboratory diagnosis of gastrointestinal infections is largely neglected in tropical African settings. This study aims to apply the Luminex multiplex PCR assay for the diagnosis of gastrointestinal pathogens in rural Ghana to evaluate its usefulness as a routine method.

Methods: A case-control study was conducted at the Agogo Presbyterian Hospital in Ghana. Stool samples were collected from children below 6 years of age with (cases) and without (controls) diarrhoea. Samples were screened for 15 different diarrhoeal pathogens by the Luminex xTAG GPP assay and associations between diarrhoea and gastrointestinal infections and fractions attributable to diarrhea (AF) were determined.

Results: The Luminex PCR assay identified organisms in 96.6 % ($n = 428$) of 443 cases and in 92.5 % ($n = 221$) of 239 selected controls. A mean of 2.5 (standard deviation [SD]: ± 1.3) and 2.3 (SD: ± 1.3) organisms per sample were detected in cases and controls respectively. An association with diarrhoea was found for rotavirus (adjusted odds ratio [aOR] = 7.2; 95 % confidence interval [CI]: 2.9–18.1), norovirus (aOR = 2.7; 95 % CI: 1.4–5.3) and *Shigella* spp. (aOR = 1.7; 95 % CI: 1.2–2.4) with respective AFs of 12.5 % (95 % CI: 9.6–15.3), 7.9 % (95 % CI: 3.8–11.7) and 16.9 % (95 % CI: 6.9–25.9).

Conclusion: The high proportion of pathogen-positive stool samples with a high number of co-infections in cases and controls suggests a substantial amount of transient or colonizing microorganisms for which treatment is not necessarily implicated. The use of sequential diagnostic algorithms with pathogen specific or quantitative PCRs might be most appropriate for diagnosing gastrointestinal infections.

Keywords: Ghana, West Africa, Multiplex Polymerase Chain Reaction, Gastrointestinal infections

What **sample size** would we need to evaluate the comparison of pathogens?

power twomeans 2.3 2.5, sd(1.3) nratio(2)

The screenshot shows the 'power twomeans' window with the following settings:

- Compute:** Total sample size (dropdown)
- Error probabilities:** 0.05 (Significance level), 0.8 (Power)
- Sample size:** 1 (Allocation ratio, N2/N1), Allow fractional sample size (unchecked)
- Effect size:**
 - Means:** 2.3 (Control), 2.5 (Experimental)
 - Standard deviations:** Common standard deviation (selected), 1.3 (Common value)
 - Group standard deviations (unchecked)
 - Assume a known standard deviations (unchecked)
- Sides:** Two-sided test (dropdown)
- Treat number lists in starred(*) options as parallel:** (unchecked)

Buttons at the bottom: Submit, Cancel, OK.

```
. power twomeans 2.3 2.5, sd(1.3) nratio(2)
```

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus Ha: m2 != m1

Study parameters:

```
alpha = 0.0500
power = 0.8000
delta = 0.2000
m1 = 2.3000
m2 = 2.5000
sd = 1.3000
N2/N1 = 2.0000
```

Estimated sample sizes:

```
N = 1,497
N1 = 499
N2 = 998
```

How much statistical **power** did we have in the comparison of pathogens?

`power twomeans 2.5 2.3, n(682) sd(1.3) nratio(2)`

power twomeans - Power analysis for a two-sample means test

Main Table Graph Iteration

Compute: Power

Error probabilities

0.05 * Significance level

Sample size

Specify: Total sample size and allocation ratio 682 * Total sample size

Allow fractional sample size 2 * Allocation ratio, N2/N1

Effect size

Means

2.3 * Control

2.5 * Experimental

Standard deviations

Common standard deviation 1.3 * Common value

Group standard deviations

* Control

* Experimental

Assume a known standard deviations

Sides: Two-sided test

Treat number lists in starred(*) options as parallel

Submit Cancel OK

```
. power twomeans 2.3 2.5, sd(1.3) n(682) nratio(2)
```

Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus Ha: m2 != m1

Study parameters:

alpha = 0.0500
N = 682
N2/N1 = 2.0000
delta = 0.2000
m1 = 2.3000
m2 = 2.5000
sd = 1.3000

Actual sample sizes:

N = 681
N1 = 227
N2 = 454

Estimated power:

power = 0.4721

nratio

```
power twomeans 2.5 2.3, sd(1.3) n(682) nratio(2)
```

- nratio is allocation ratio
- No need to specify if it is 1 (equal number in each group)
- This example data comes from a case control study where there were approximately 2:1 cases to controls

Visualizing power/sample size curves

power twomeans - Power analysis for a two-sample means test

Main Table **Graph** Iteration

Compute: Power * Accepts numlist [\(Examples\)](#)

Error probabilities
0.05 * Significance level

Sample size

Specify:
Total sample size and allocation ratio 682 * Total sample size
☐ Allow fractional sample size 2 * Allocation ratio, N2/N1

Effect size

Means
2.3 * Control
2.5 * Difference

Standard deviations
☒ Common standard deviation
1 * Common value
☐ Group standard deviations
 * Control
 * Experimental
☐ Assume a known standard deviations

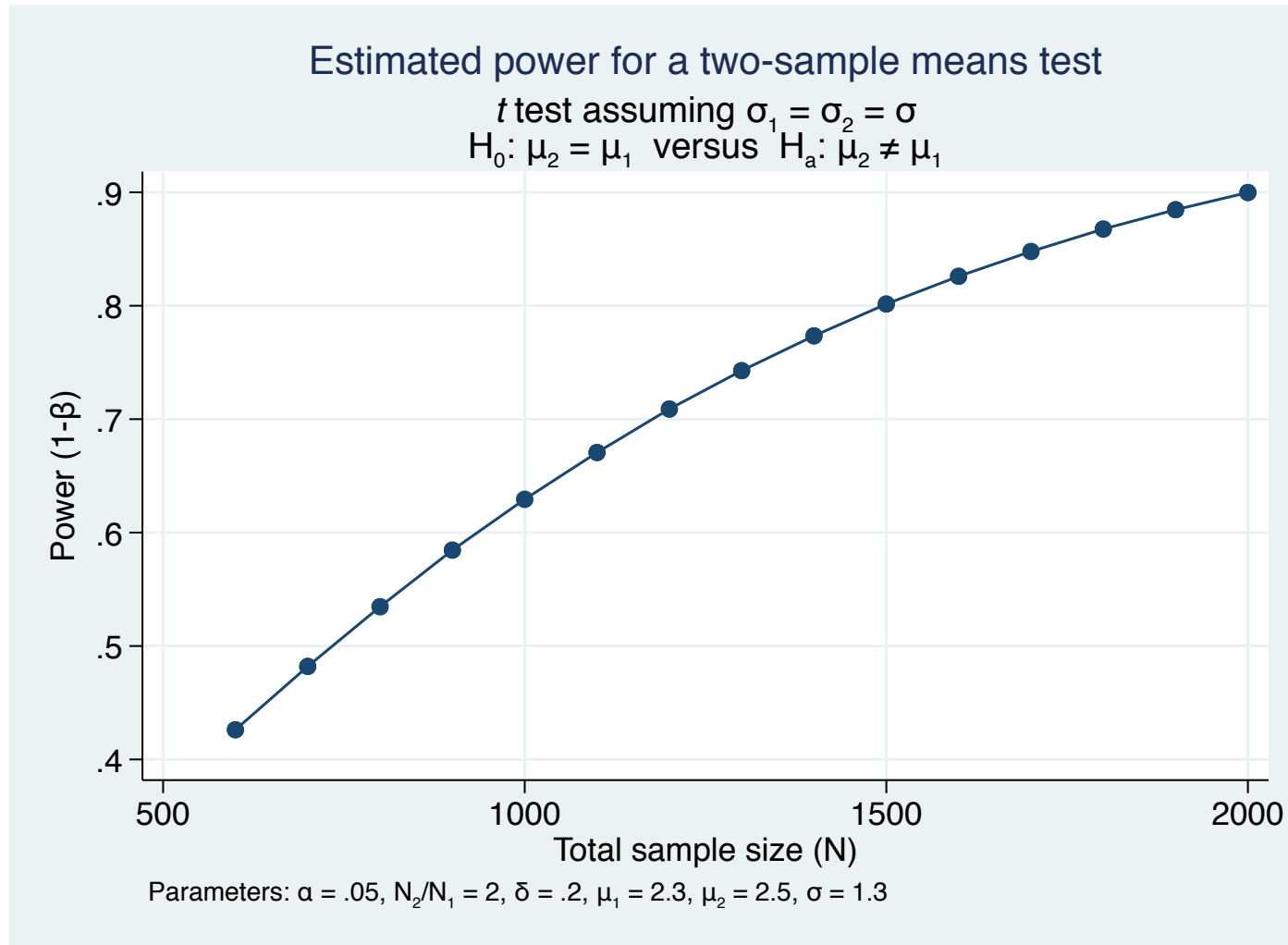
Sides:
Two-sided test

☐ Treat number lists in starred(*) options as parallel

? R Submit Cancel OK

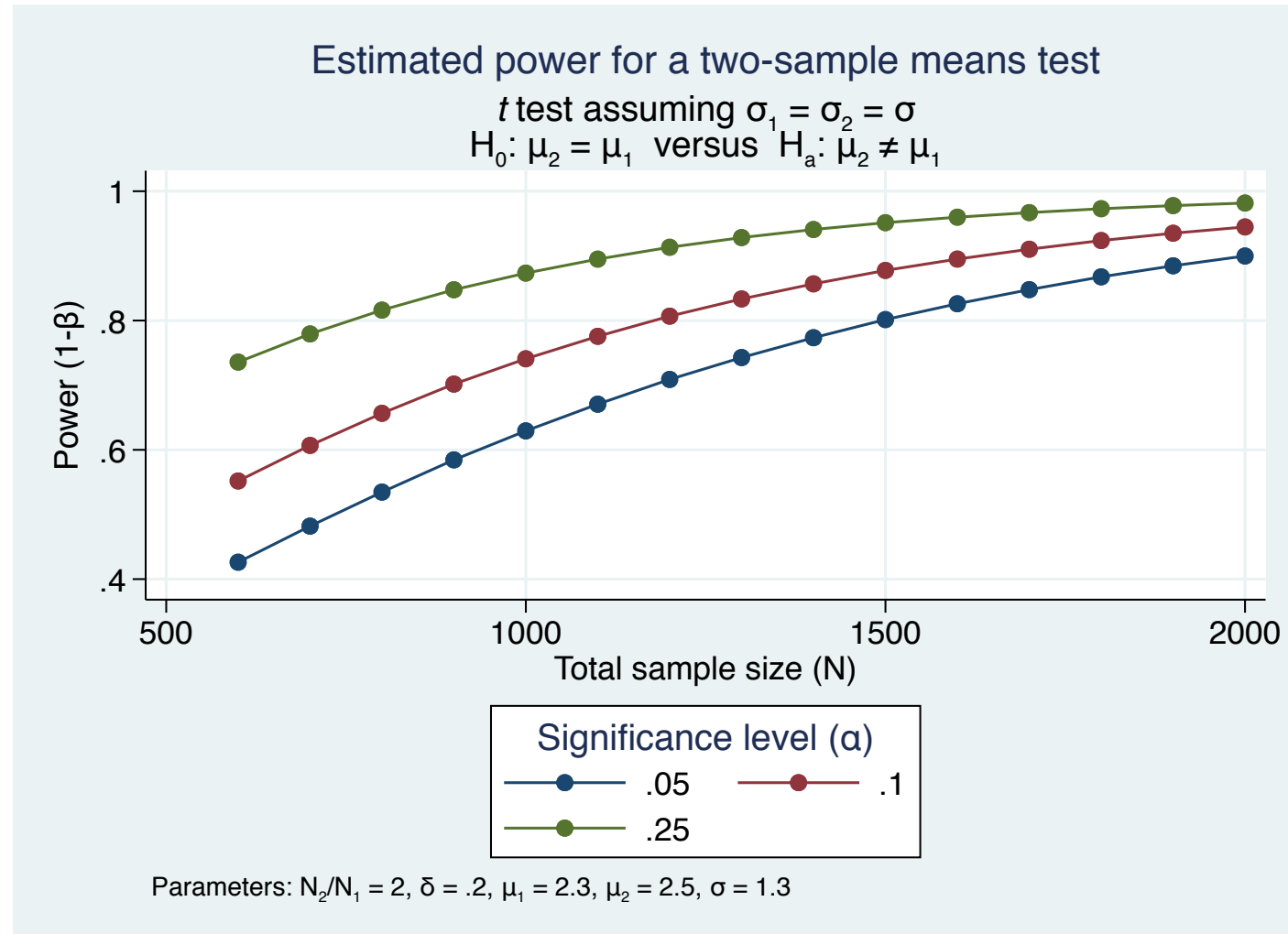
Visualizing power/sample size curves

```
power twomeans 2.3 2.5, n(600(100)2000) sd(1.3) nratio(2) graph
```



Visualizing power/sample size curves

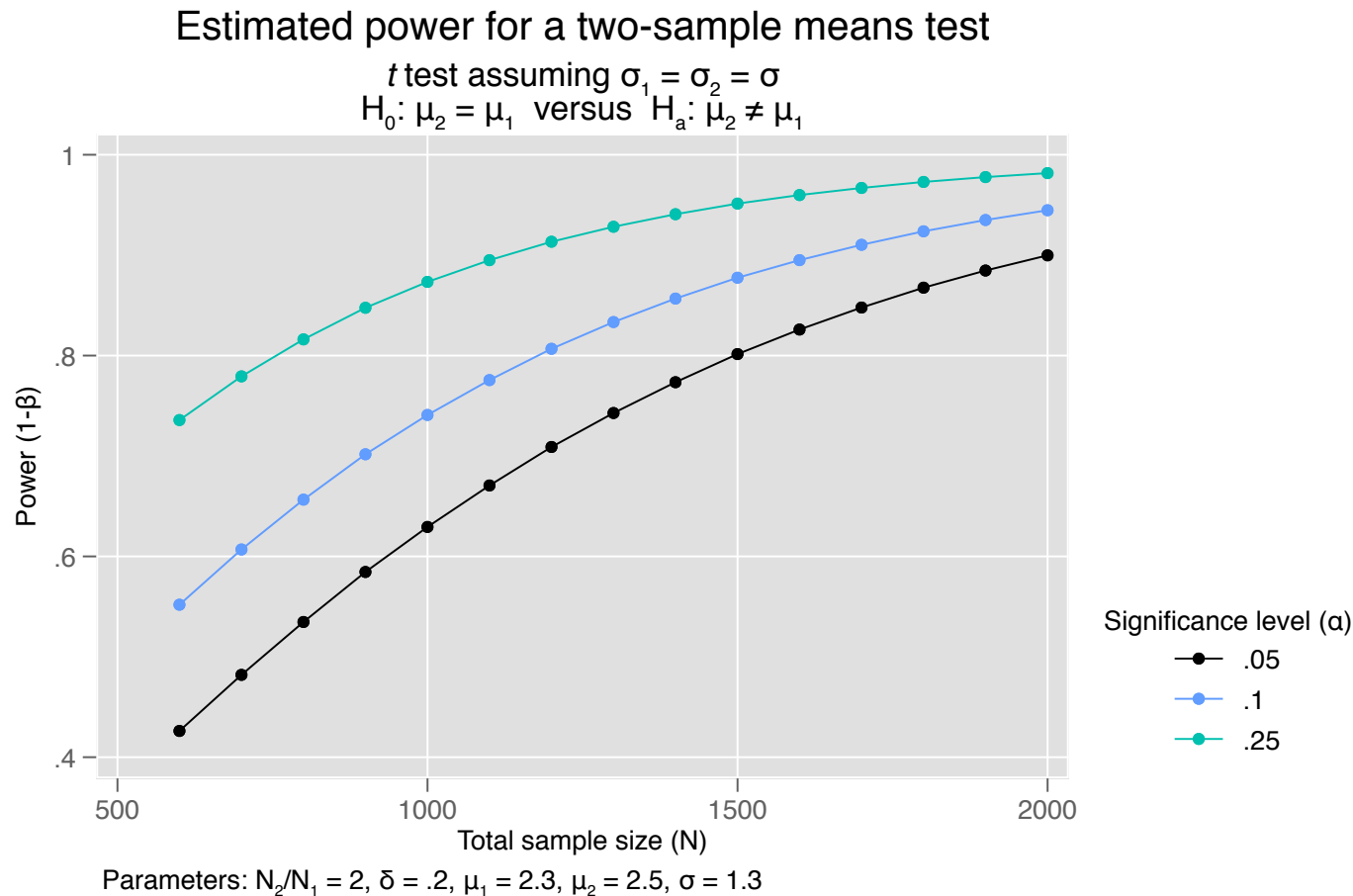
```
power twomeans 2.3 2.5, n(600(100)2000) alpha(.05 .10 .25) sd(1.3) nratio(2) graph
```



Visualizing power/sample size curves

set scheme plottig

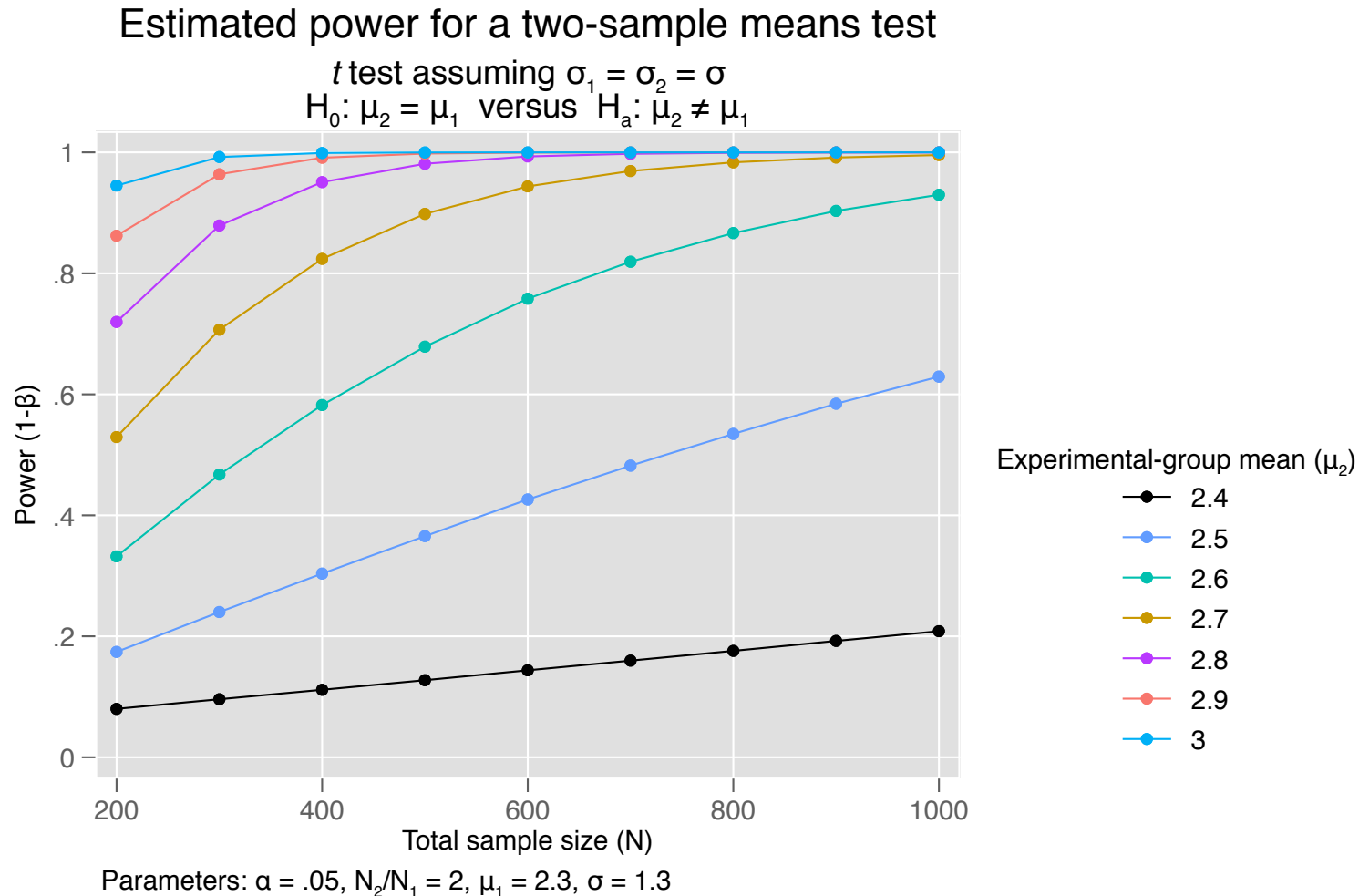
```
power twomeans 2.3 2.5, n(200(50)1000) sd(1.3) alpha(.05 .10 .25) nratio(2) graph
```



*Need to download this scheme first:
ssc install blindschemes

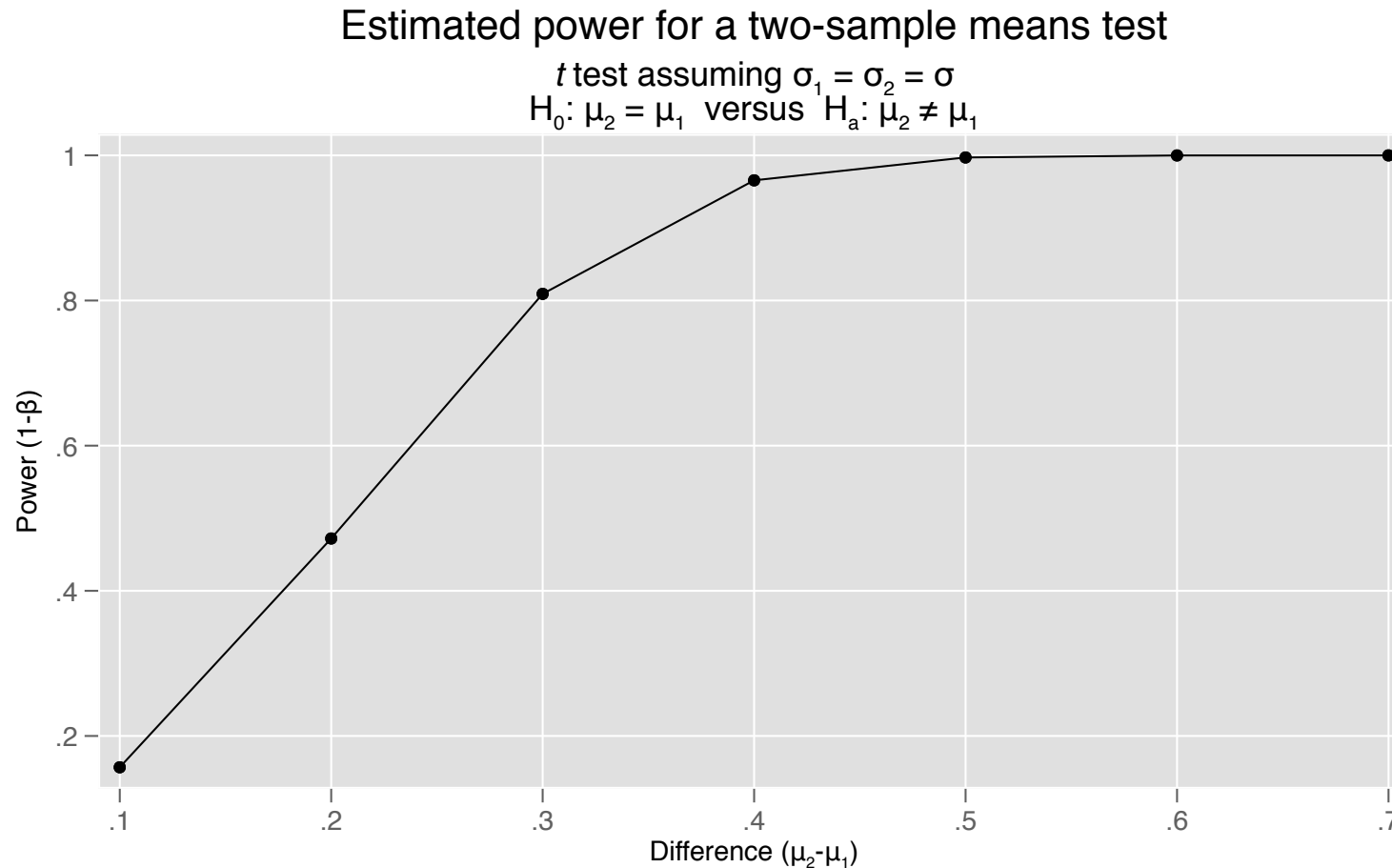
Visualizing power/sample size curves

```
power twomeans 2.3 (2.4(.1)3), n(200(100)1000) sd(1.3) nratio(2) graph
```



Visualizing power/sample size curves

```
power twomeans 2.3, diff(.1(.1).7) n(682) sd(1.3) nratio(2) graph
```



Visualizing power/sample size curves

```
power twomeans 2.3, diff(.1(.1).7) n(200(50)1000) sd(1.3) nratio(2) graph
```

