Two Sample Comparison and Bivariate Regression

Aaron McAdie

Topics

- Replicates vs Repeated Measures
- 2 Sample Comparisons
  - simulation examples of S/N, α, power
- Bivariate Linear Fits
  - statistical underpinnings
  - model evaluation (R^2, residuals, confidence intervals)
  - explanation vs prediction

Replication vs Repeated Measures

- Replication – completely duplicating measurement with new samples
  - i.e. measuring 5 random people's height
- Repeated measures – measuring the same sample multiple times
  - i.e. measuring one person's height 5 times
- Replication captures sample variability and allows inference about a population, repeated measures capture test variability

Two Sample Comparisons

- Even with multiple data points it is easy to draw incorrect conclusions about whether or not samples are the same/similar
- t-tests can protect against Type I error (false discovery) to a degree
- increasing sample size decreases Type II error rate (failing to find an effect that is present)
- simulations where underlying population parameters are known help to illustrate these concepts

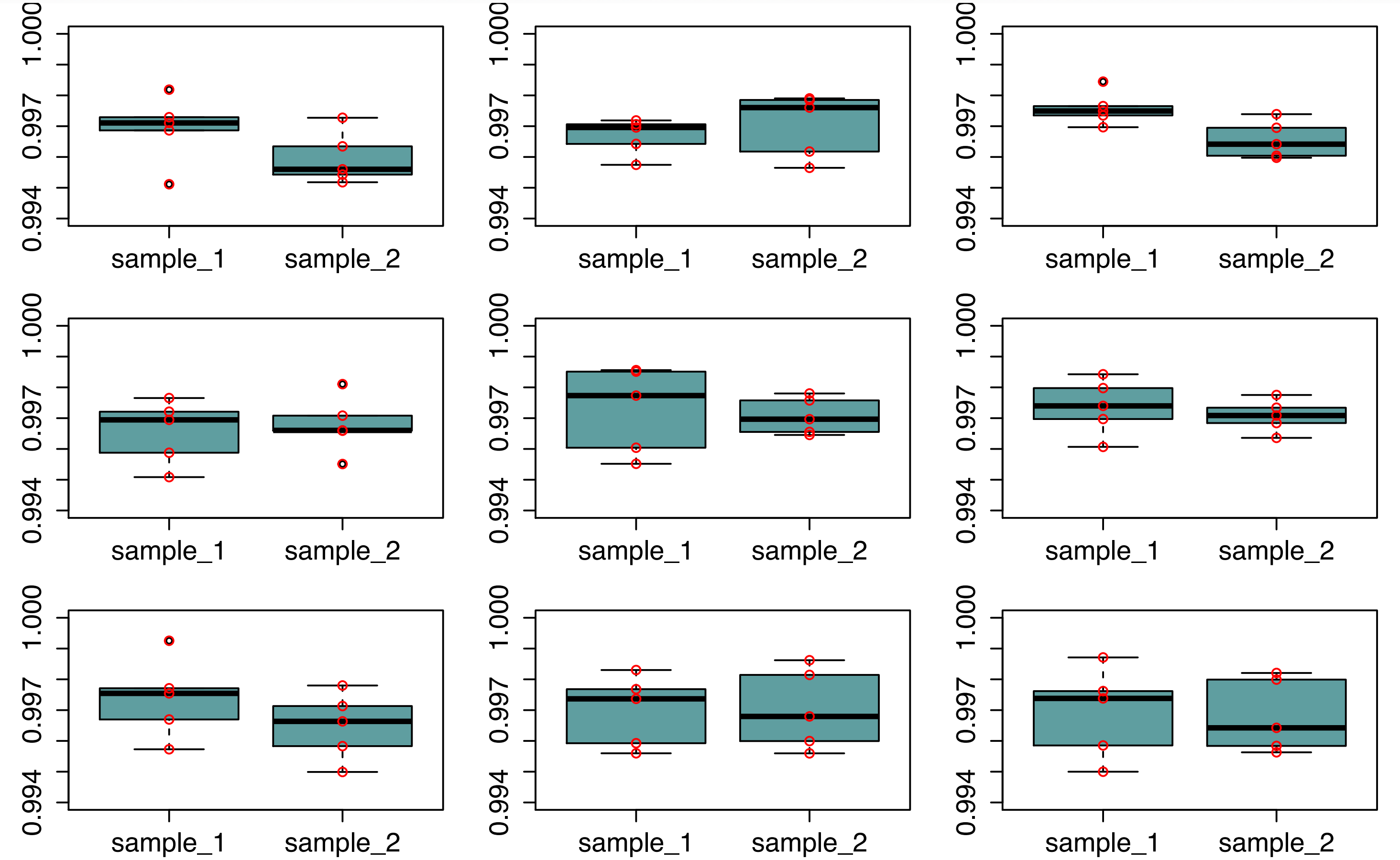
Two Sample Comparisons

- Simulation 1
  - 5 random samples drawn from two populations
  - population 1 ~ N(μ = 0.997, σ = 0.001, σ^2 = 1e-6)
  - population 2 ~ N(0.997, 1e-6)
- t test is performed on two samples, pvals recorded
- new set of random draws is generated, pvals recorded
- simulation run 1000 times
- proportion of $p<0.05$ calculated

Two Sample Comparisons

Simulation 1 (equal means, n=5)

-proportion of p<0.05 = 0.051



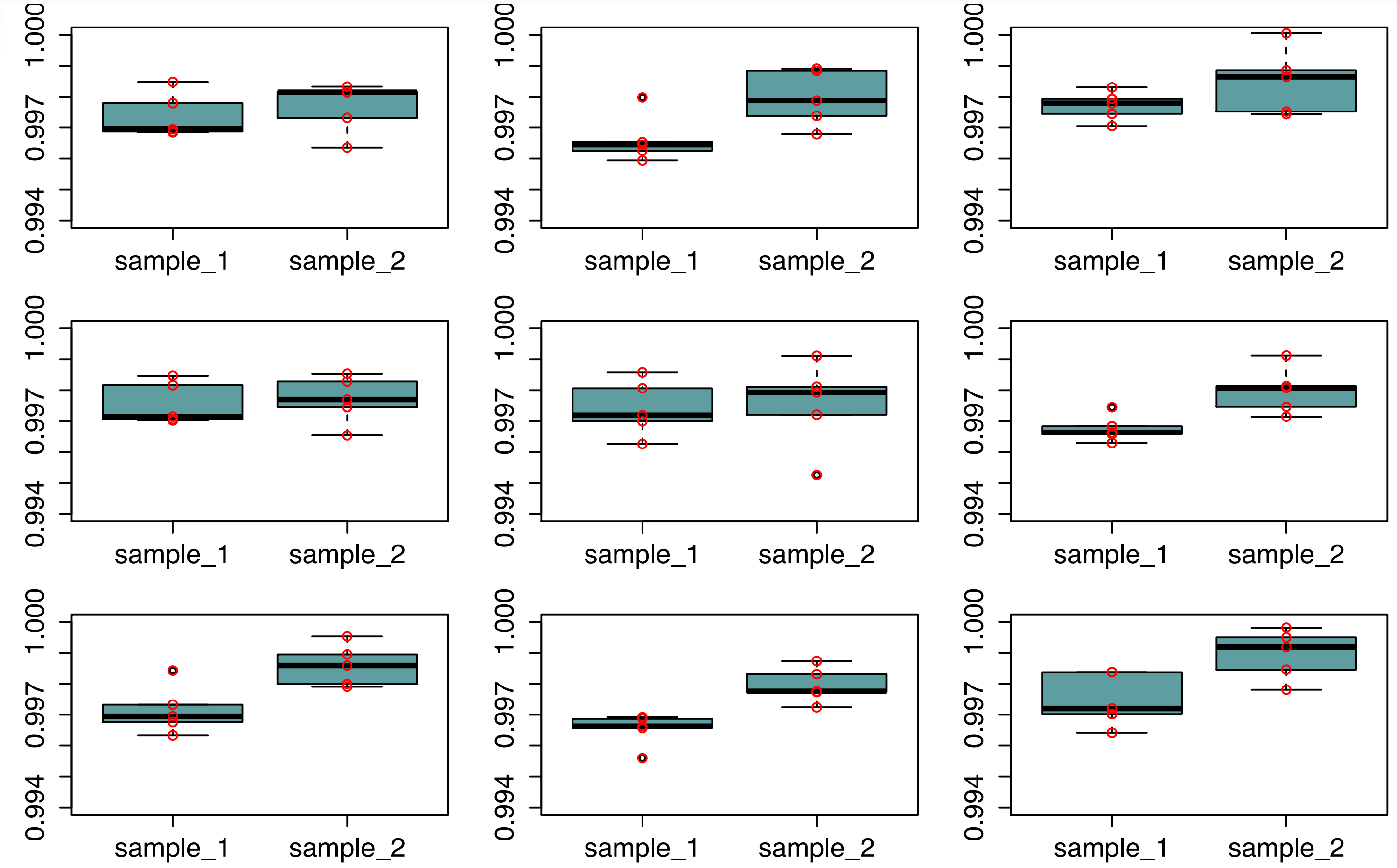
Two Sample Comparisons

- Simulation 2
  - 5 random samples drawn from two populations
  - population 1 ~ $N(0.997, 1e\text{-}6)$
  - population 2 ~ $N(0.998, 1e\text{-}6)$

Two Sample Comparisons

Simulation 2 (unequal means, n=5)

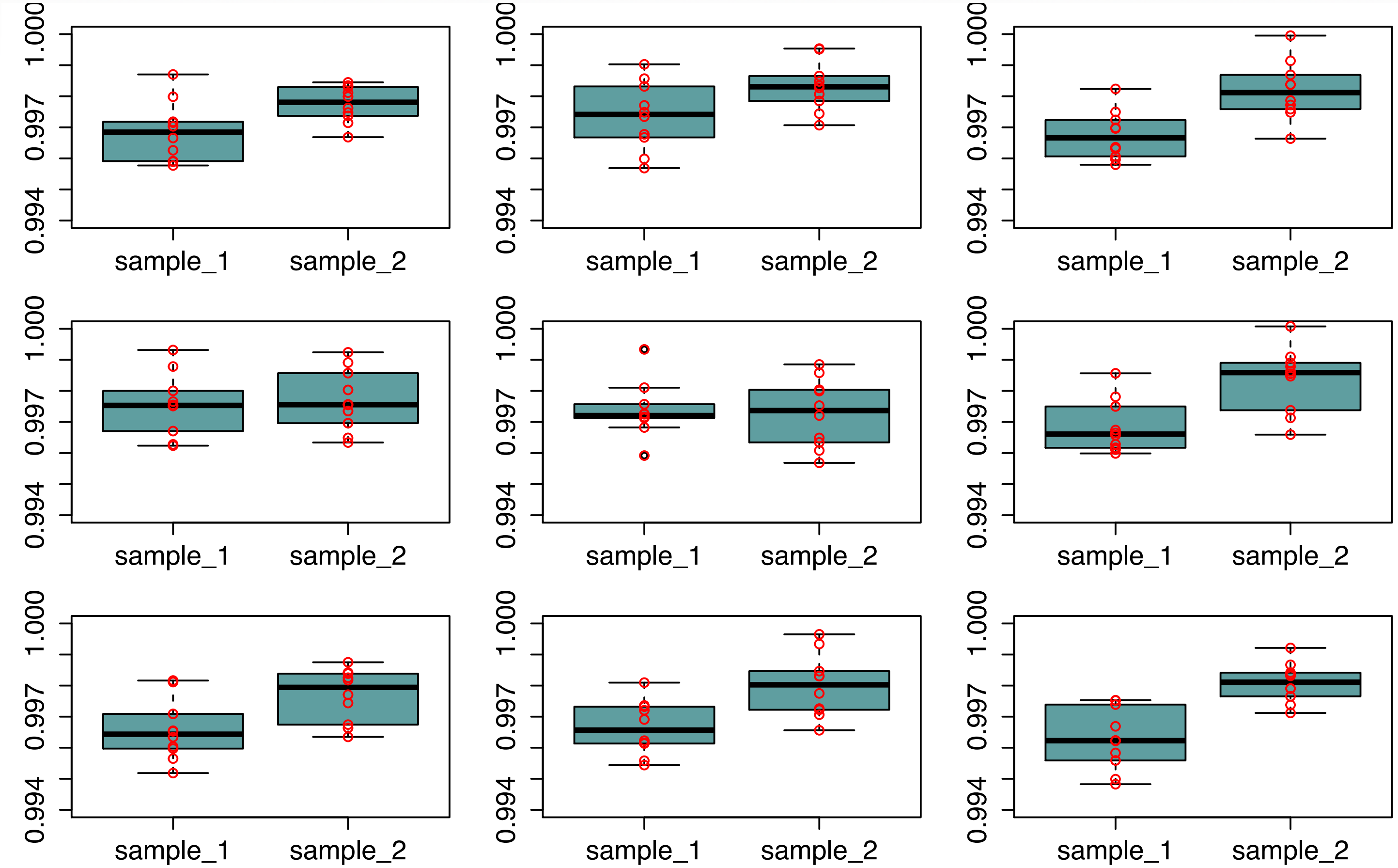
-proportion of $p<0.05$ = 0.281



Two Sample Comparisons

Simulation 3 (unequal means, n=10)

-proportion of p<0.05 = 0.554



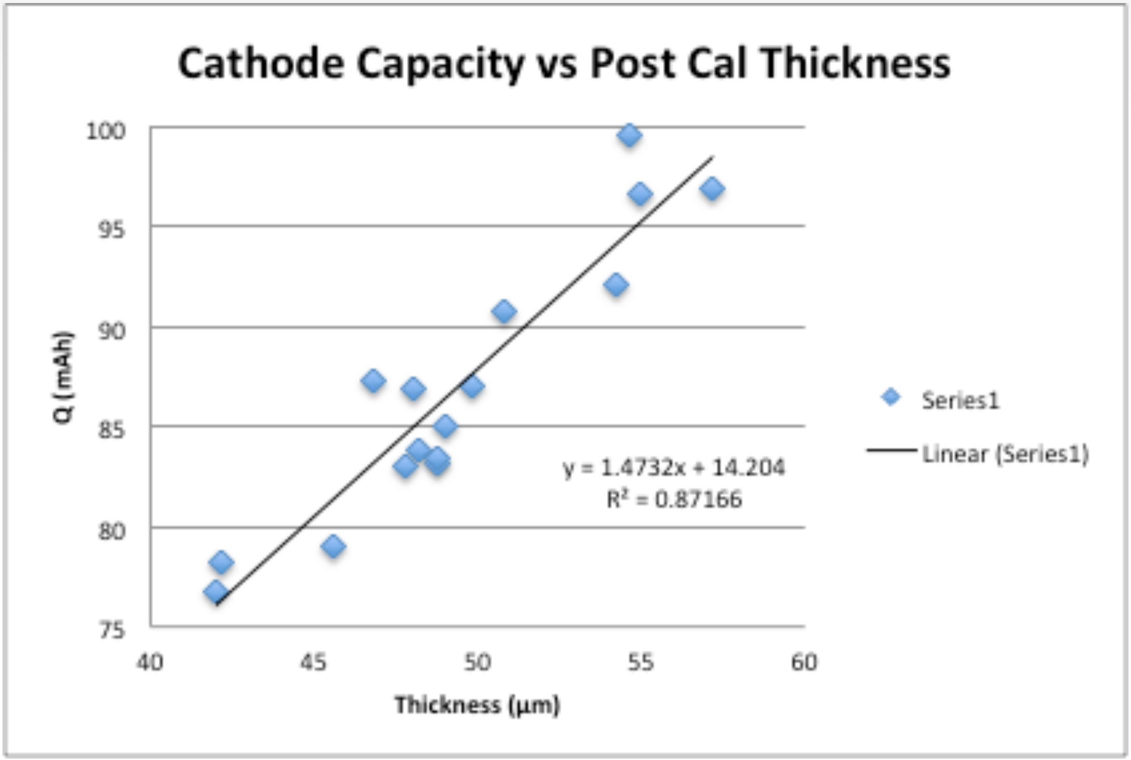
Sample Size and Power

- Power is a function of μ_1-μ_2, σ, n, α
- table at right applies for μ_1=0.997, μ_2=0.998, σ=0.001, α=0.05

n	Power
5	0.281
10	0.554
20	0.864
30	0.975

- simulations use data pulled from normal distributions and with common variance. Methods exist for unequal variances and non-parametric data.

Linear Least Squares Regression



Linear Least Squares Regression

- The line that minimizes the sum of the squared distances between each observation and the line

$$\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_1))^2$$

- Estimate of slope term -> $\hat{\beta}_1 = cor(Y, X)\dfrac{sd(Y)}{sd(X)}$

- Estimate of intercept term -> $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

- Regression Model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

R^2 and Beyond

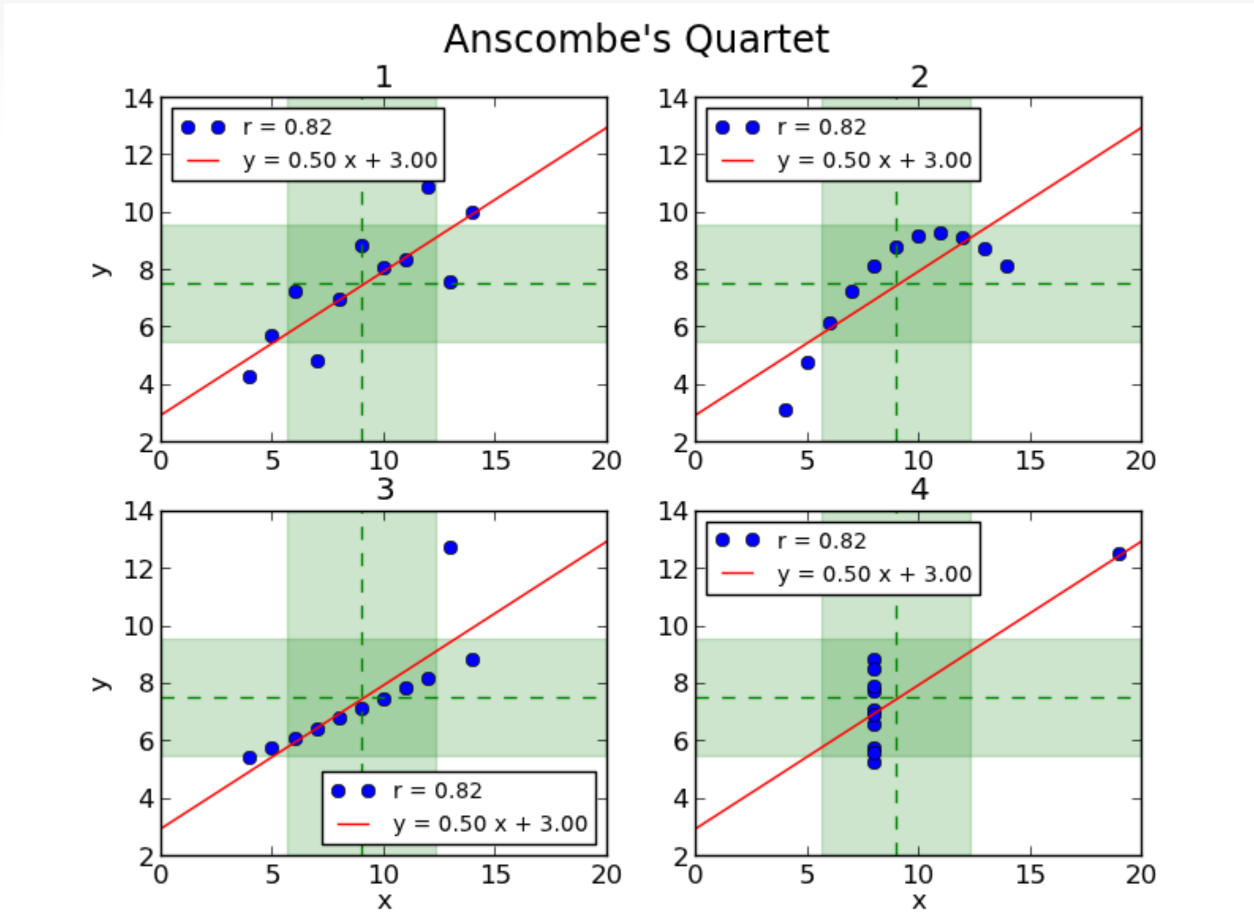
- R^2 is the percentage of variability in Y explained by the regression model

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2}$$

- R^2 provides useful information but can be a misleading statistic
- it is part of the story, but not the whole story

R^2 and Beyond

R^2 can be a misleading statistic



source: http://informatique-python.readthedocs.org/fr/latest/Exercices/anscombe.html

R^2 and Beyond

- Tools to assess uncertainty and model fit
  - standard error of estimates
  - confidence intervals
  - residual diagnostics

Uncertainty of Parameter Estimates

- Standard error is a measure of estimate precision. It is the standard deviation of the *sampling distribution*
- standard error of regression slope:

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}$$

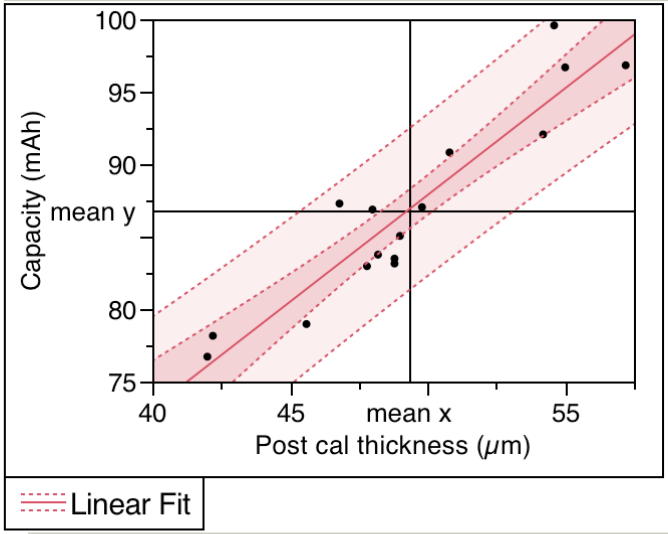
- allows for determination of practical significance vs statistical significance

Confidence and Prediction Intervals

- A confidence interval around a fitted line represents uncertainty in model fit (parameter estimation)
- A prediction interval around a fitted line represents uncertainty in predicting a new y_i given x_i
- Both confidence and prediction intervals are the smallest in the center of the data
- Prediction intervals are always wider than confidence intervals

Confidence and Prediction Intervals

Bivariate Fit of Capacity (mAh) By Post cal thickness (μm)



$$\sigma_{fit,x_o} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}$$

$$\sigma_{pred,x_o} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}$$

Linear Fit

Capacity (mAh) = 14.204149 + 1.4732272*Post cal thickness (μm)

Summary of Fit

RSquare	0.871658
RSquare Adj	0.862491
Root Mean Square Error	2.523045
Mean of Response	86.83425
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>ltl
Intercept	14.204149	7.475063	1.90	0.0782
Post cal thickness (μm)	1.4732272	0.151083	9.75	<.0001*

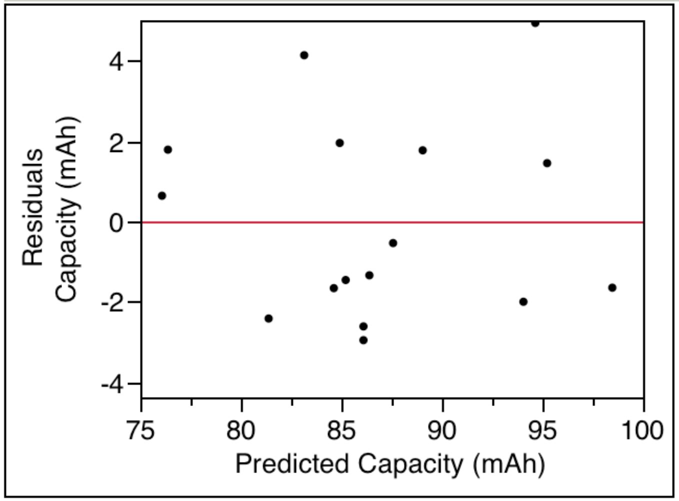
Residual Diagnostics

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \boxed{\varepsilon \qquad \varepsilon \sim N(0, \sigma^2)}$$

- Residual diagnostics are an important part of model validation
- Check to see that residuals are more or less normally distributed and look for patterns across your modeling space

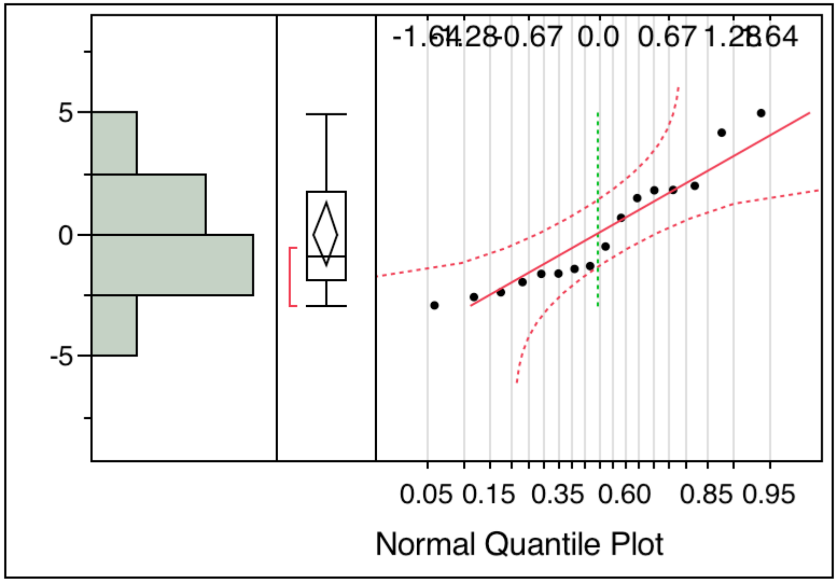
Residual Diagnostics

Bivariate Fit of Residuals Capacity (mAh) By Predicted Capacity (mAh)

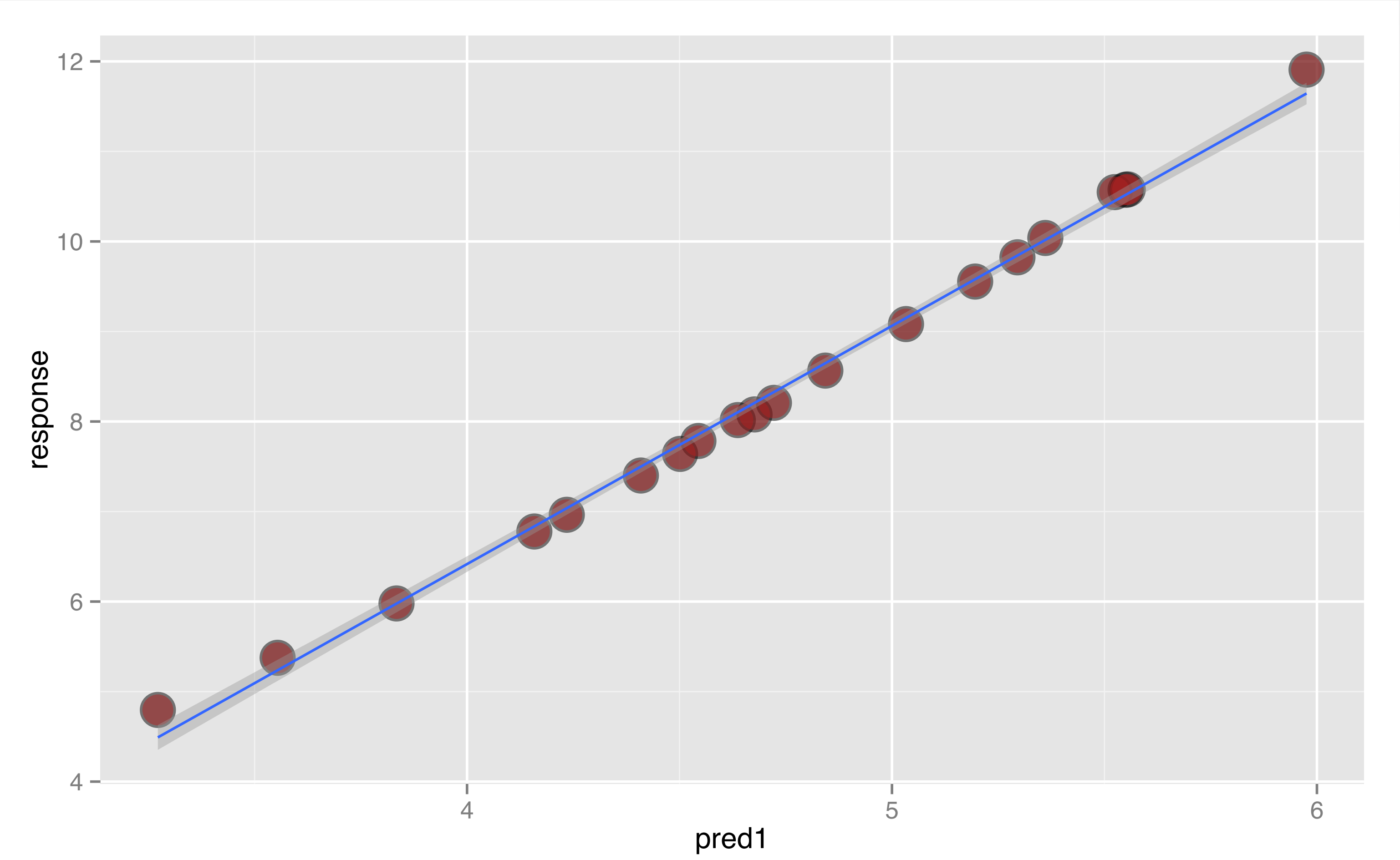


Distributions

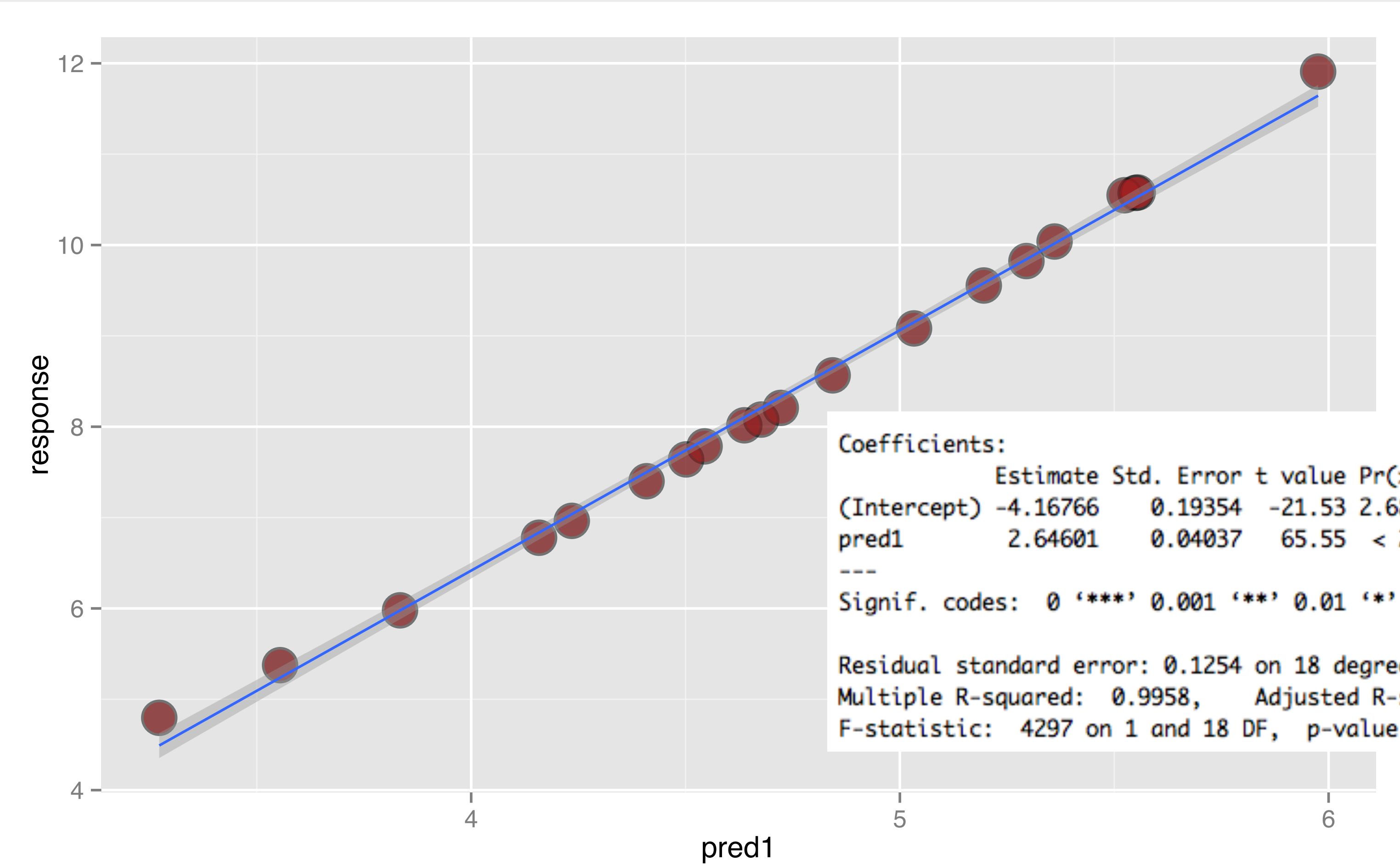
Residuals Capacity (mAh)



Residual Diagnostics



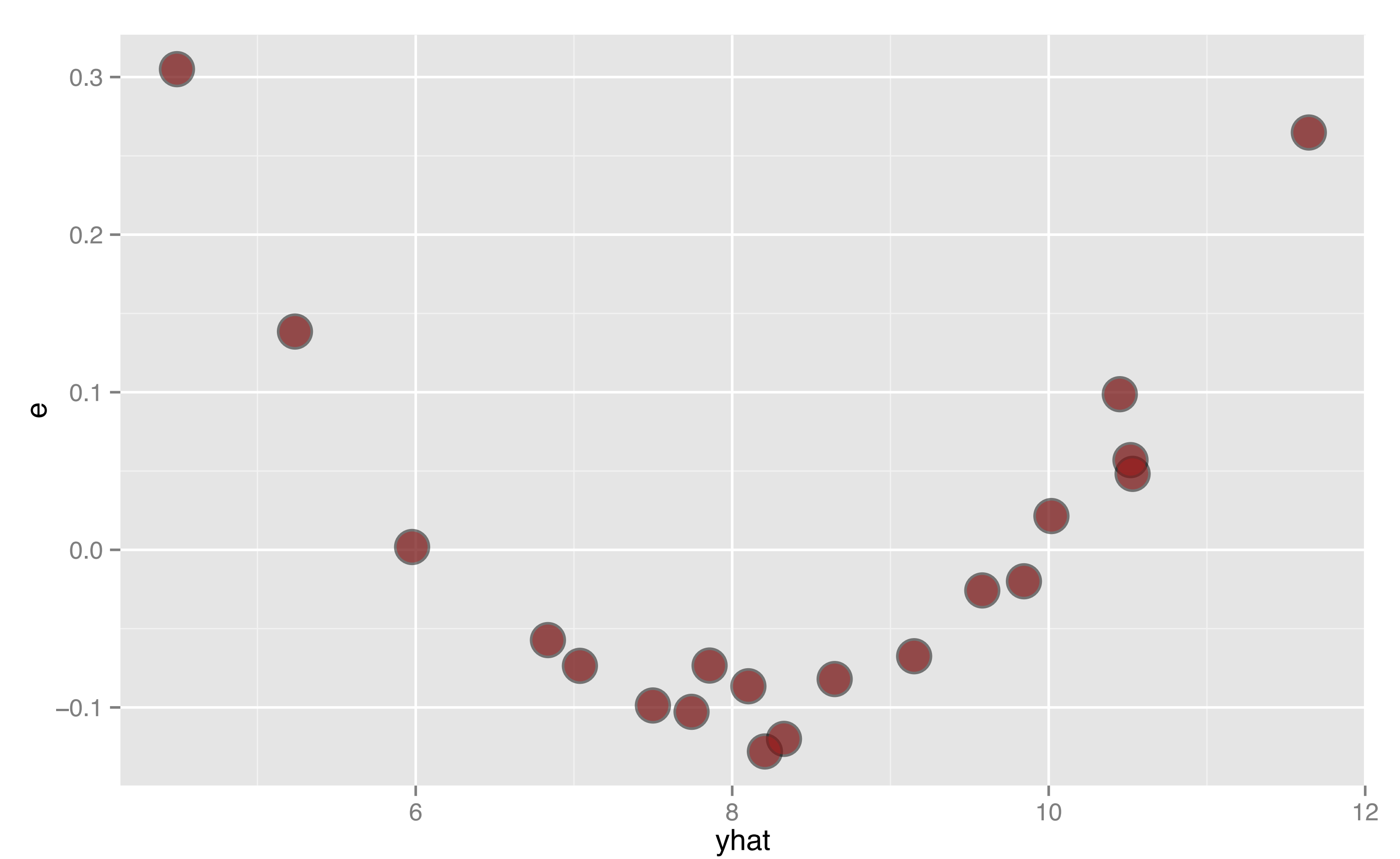
Residual Diagnostics




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.16766    0.19354  -21.53 2.68e-14 ***
pred1        2.64601    0.04037   65.55  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1254 on 18 degrees of freedom
Multiple R-squared:  0.9958,    Adjusted R-squared:  0.9956
F-statistic:  4297 on 1 and 18 DF,  p-value: < 2.2e-16
```

Residual Diagnostics



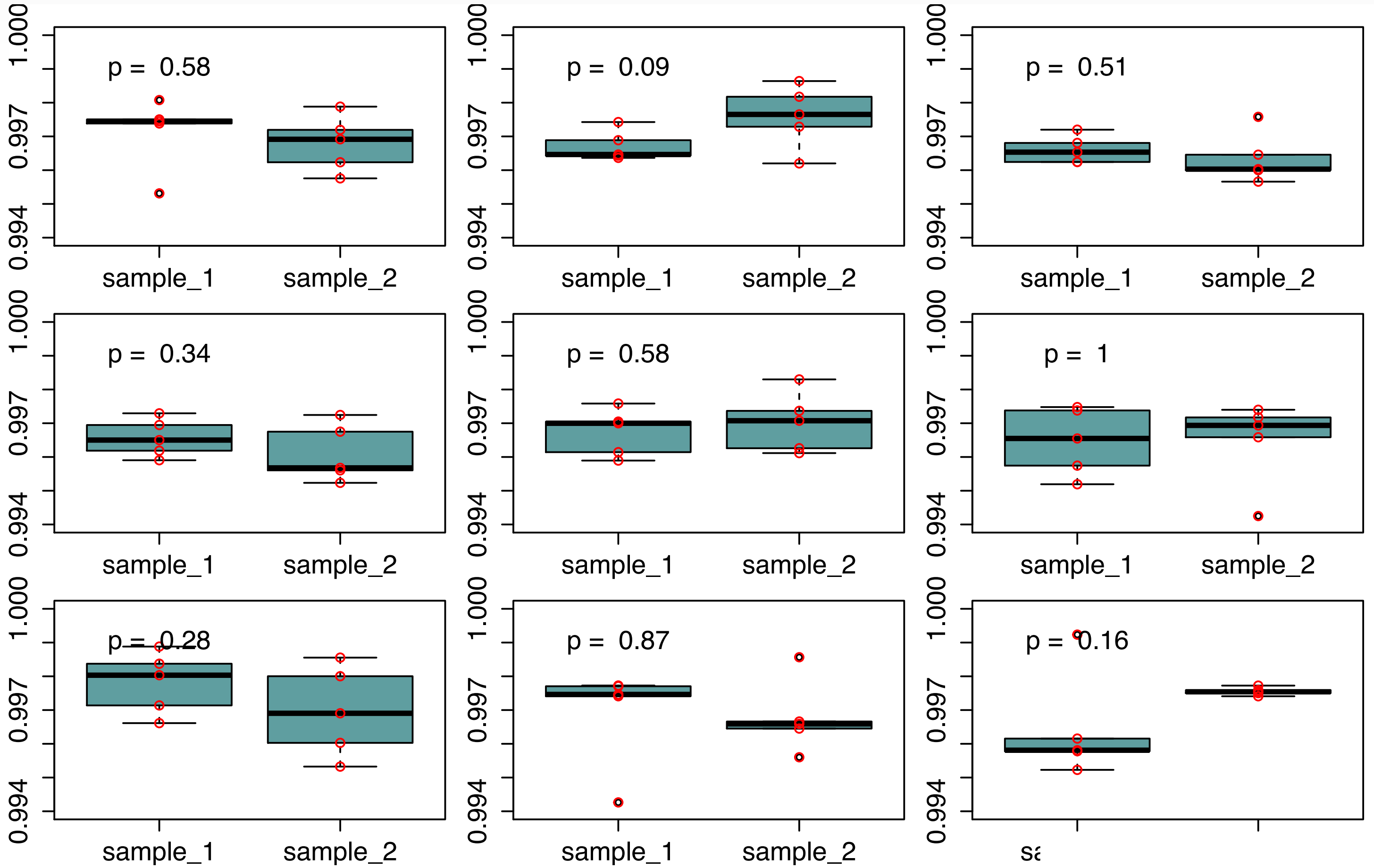
Follow-Up Questions

- What does it look like if p-values are overlaid on simulation boxplots?
- Does decreasing test error affect statistical power more than increasing sample size?
- Is it better to run n=2 in 4 tests or n=8 in one test?
- What is the interpretation of prediction intervals? If the y_{i+1} observation falls within the prediction interval, does that mean the regression model is correct?

Equal Means Simulation with P-values

Simulation 4 (μ_1=0.997, μ_2=0.997, σ=0.001, n=5)

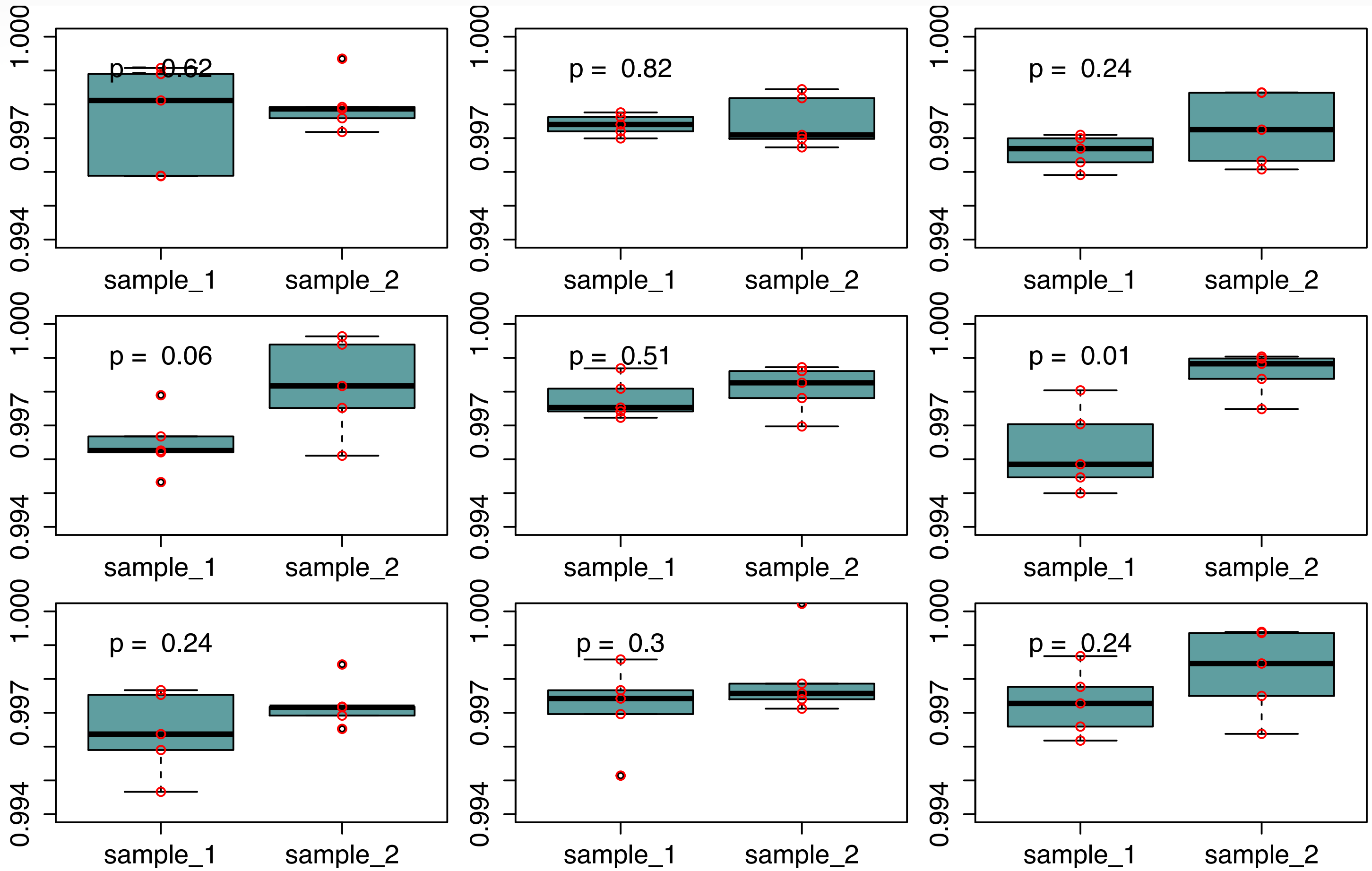
-proportion of $p<0.05$ = 0.044



Population Variability, Sample Size, and Power

Simulation 4 (μ_1=0.997, μ_2=0.998, σ=0.001, n=5)

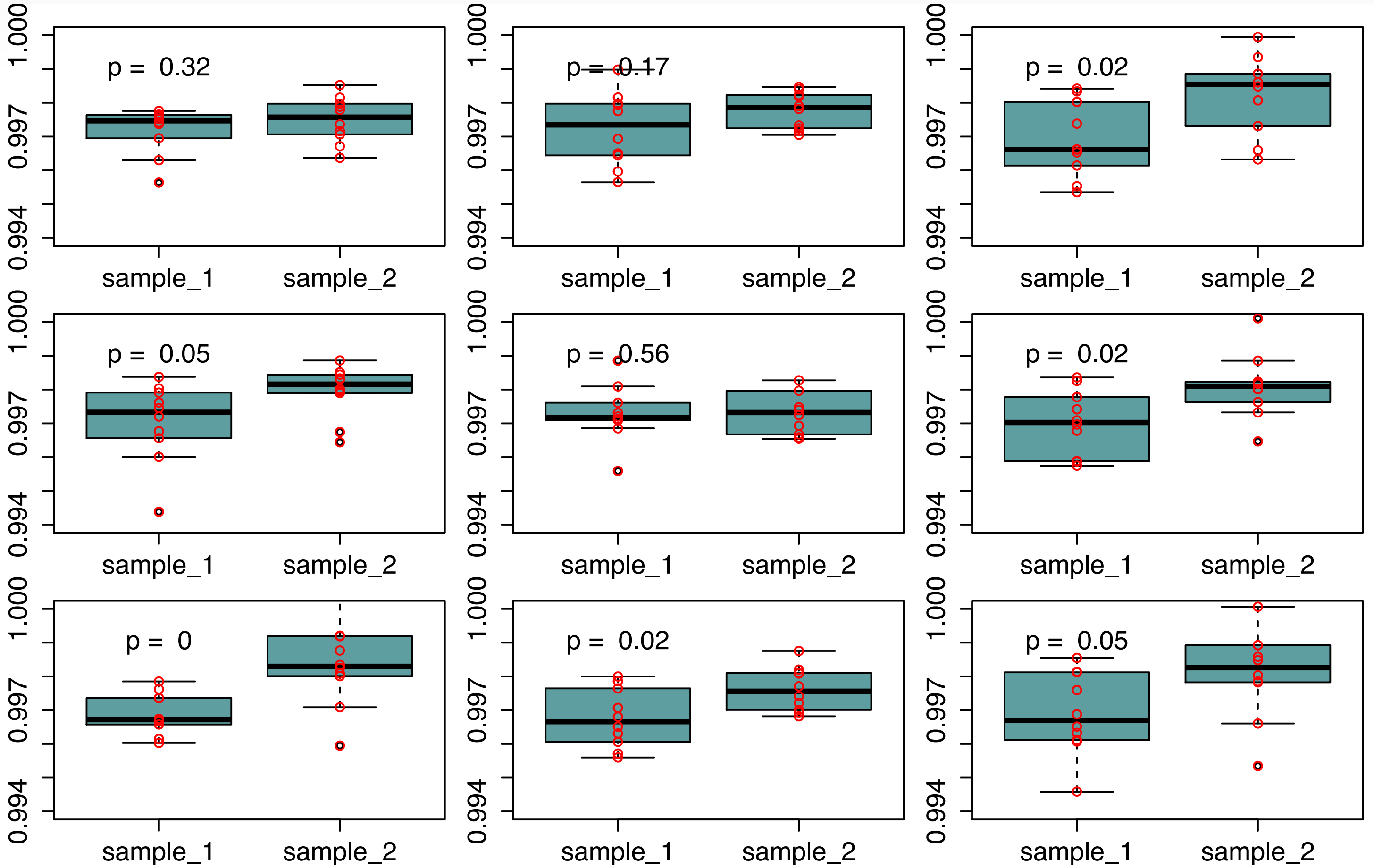
-proportion of $p<0.05$ = 0.287



Population Variability, Sample Size, and Power

Simulation 5 (μ_1=0.997, μ_2=0.998, σ=0.001, n=10)

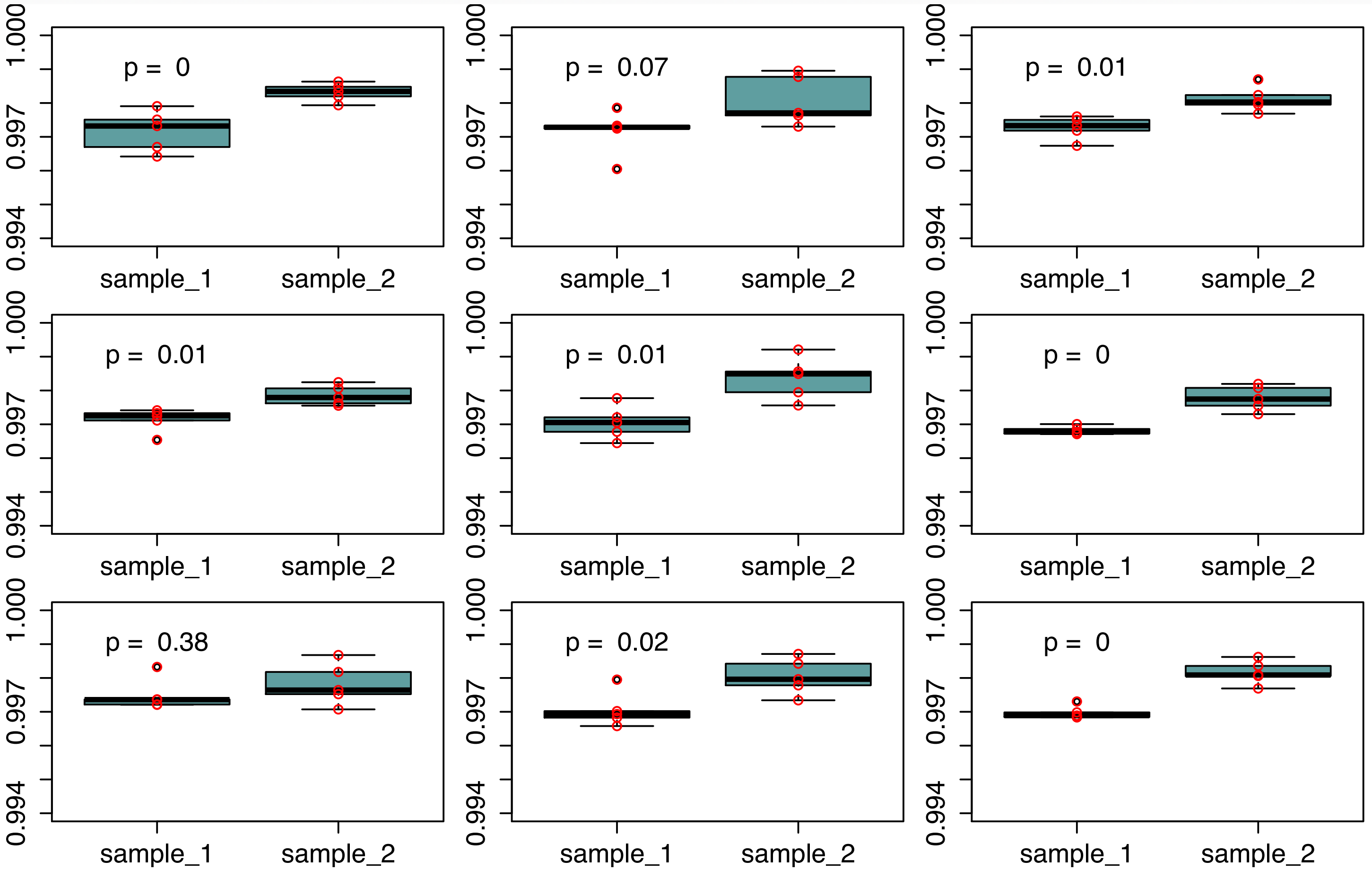
-proportion of $p<0.05$ = 0.56



Population Variability, Sample Size, and Power

Simulation 5 (μ_1=0.997, μ_2=0.998, σ=0.0005, n=5)

-proportion of p<0.05 = 0.784



Population Variability, Sample Size, and Power

- Halving variability has a larger effect on power than doubling sample size
- Power is the probability of rejecting a null hypothesis (H_o: $\mu_1 = \mu_2$) when it is false

$$P\left(\frac{\mu_1 - \mu_2}{\sigma/\sqrt{n}} > t_{1-\alpha,n-1}; \mu_1 \neq \mu_2\right)$$

*assuming pooled variance and $n_1 = n_2$

- Power is proportional to $1/\sigma$ and $\sim n^{1/2}$
  - not directly proportional to $n^{1/2}$ because the critical t value also changes with n, although not much

Spreading Samples Across Multiple Tests

- **Is it better to run n=2 in 4 similar tests or n=8 in one test?**
- Statistically speaking running 4 tests with n=2 would quadruple the risk of a false discovery, and also give lower power for each individual test, so it might result in 8 samples of inconclusive or misleading data
- If tests were different enough to justify, it could be possible to create a combined metric using weighting factors and then do a non-parametric test such as a Wilcoxon Rank Sum, however I've never seen this done so it might not be valid

Prediction Interval Interpretations

- Prediction intervals, like confidence intervals have a somewhat tricky interpretation
- if you collect n xy pairs from a population, construct a regression line and a 95% prediction interval, then collect an observation n+1, and do that over and over, that n+1 observation will fall within the prediction interval 95% of the time, on average
- *not* representative of the distribution of future observations at a given x value
- Prediction intervals are the safest when considered a plausible range for new data