Multiple Testing & Permutations

Lauren Vanderlinden, PhD, MS
T15 Postdoctoral Fellow Computational Biology
Division of Rheumatology & Department of Biomedical Informatics
School of Medicine, University of Colorado Anschutz Medical Campus

CPBS 7602 - December 10, 2024

Outline

- Multiple testing burden
- Inflation factor
- Methods to adjust for multiple testing
- Permutations
- Hands on activity

What is Multiple Testing & Why It Is A Problem

Multiple Testing

- A single statistical tests carries a 5% risk of producing a false positive conclusion
  - Generally considered an acceptable level of risk

- Repeated tests rack up a much greater (and ultimately unacceptable) risk.

Why is multiple testing a problem?

- If we test 20 hypotheses where there is truly no difference, we expect that at one test will have ‘significant’ p-value ($p<0.05$), e.g., a false positive.
- Likewise, the probability of at least one false positive if we perform even 5 tests is:

$$\begin{aligned}\text{Probability of at least one mistake} &= 1 - \text{Probability of no mistakes} \\ &= 1 - Pr(\text{no mistake})^5 \\ &= 1 - (0.95)^5 \\ &= 1 - 0.77 \\ &= 0.22\end{aligned}$$

Why is multiple testing a problem?

- Taken to extremes, multiple testing is virtually guaranteed to find 'statistical significance' even in the absence of any real effects.
- Probability of at least one mistake out of 100 tests:

$$\begin{aligned}\text{Probability of at least one mistake} &= 1 - \text{Probability of no mistakes} \\ &= 1 - Pr(\text{no mistake})^{100} \\ &= 1 - (0.95)^{100} \\ &= 1 - 0.0059 \\ &= 0.9941\end{aligned}$$

YIKES!

Where does multiple testing arise?

- Comparing several treatment groups
- Recording numerous endpoints and testing each one for changes
- Measuring the same end point on several occasions and testing at each time point
- Breaking the data into numerous sub-sets and testing within each of these
- Examples: GWAS, Image analyses (voxel-wide testing), survey data, etc.

Comparing multiple treatment groups

- If we assess several possible treatments, it is tempting to make all pairwise comparisons.
- As the number of groups increases, the number of comparisons increases dramatically, e.g.,
  - 3 groups = 3 comparisons
  - 4 groups = 6 comparisons
  - 10 groups = 45 comparisons

$$n\ groups = \binom{n}{2} = \frac{n!}{2!\,(n-2)!} = \frac{n(n-1)}{2}$$

Comparing several endpoints

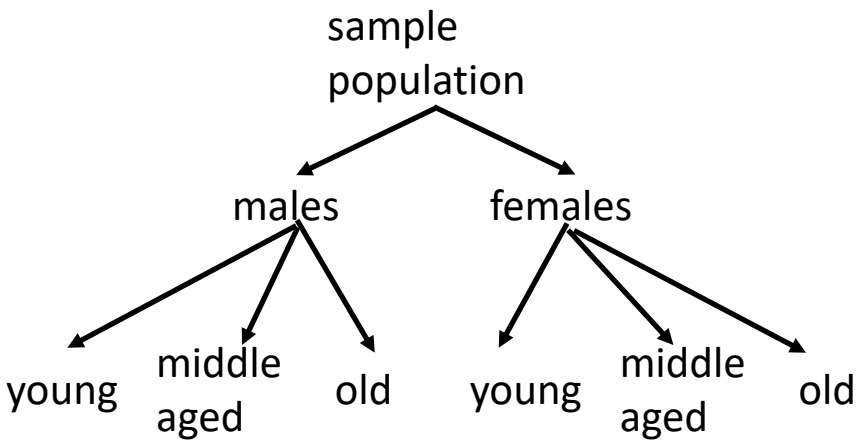
- It is tempting to measure lots of variables in a study and testing each outcome compared to our intervention.
- This is especially true in omics studies where the number of endpoints can range from a few hundred to over a hundred thousand.
- This also applies to multiple predictors of the same outcome, e.g., genome-wide association studies

Testing several timepoints

- The relevant endpoint is often measured at several time points, and it is tempting to test for differences as each time point.

Testing within numerous groups

- It is tempting when results are not statistically significant in the total sample population to begin to break the sample populations into subgroups and check for statistical significance.
- But is becomes a slippery slope…

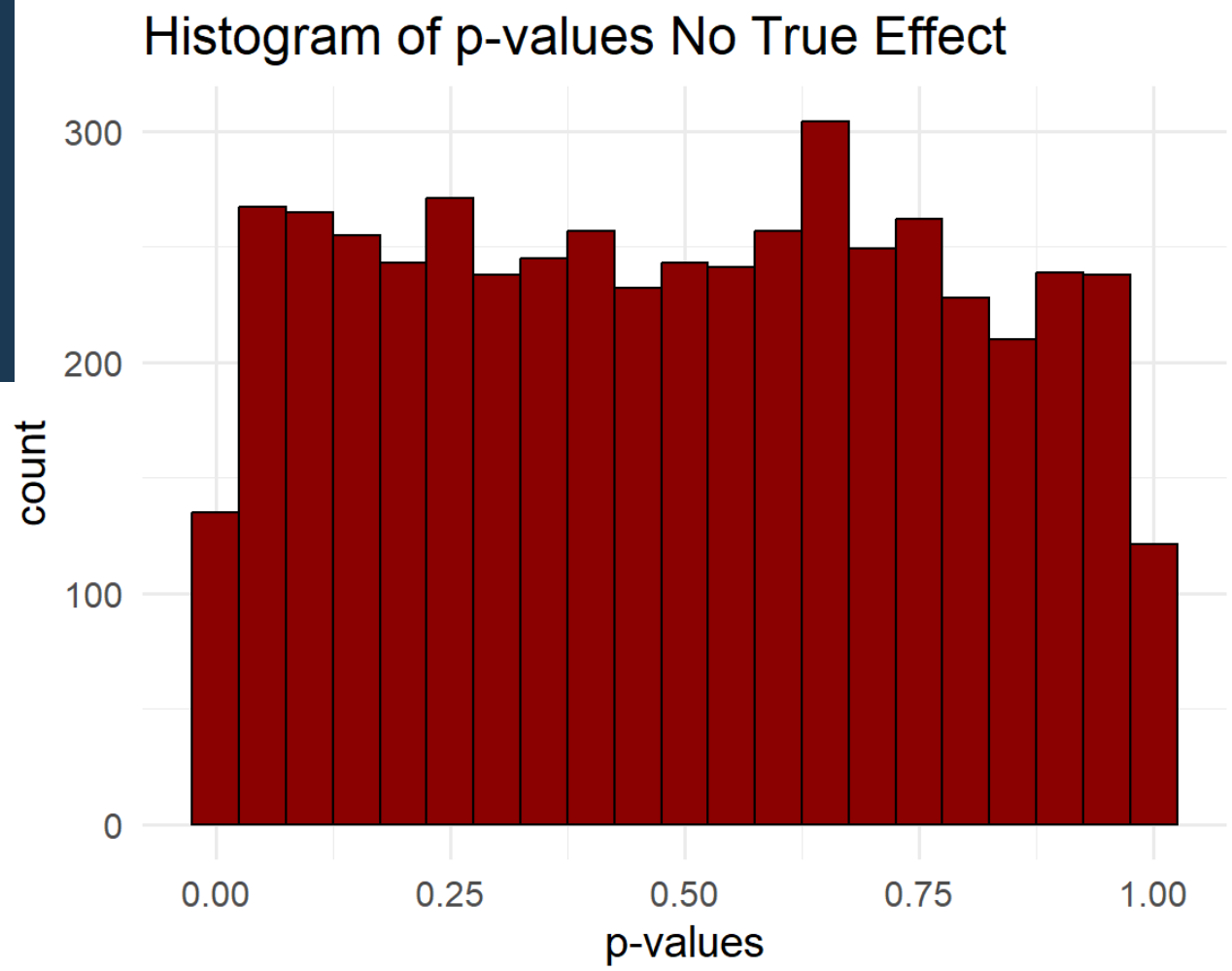


Expected distribution of p-values in multiple testing scenarios

Distribution of p-values when no true effect

```r
# simulated data for a 2-sample t-test
ngenes <- 5000
nsamples <- 100

pvals <- apply(as.matrix(1:ngenes),
               1, function(a)
                 t.test(rnorm(n = nsamples),
                   rnorm(n = nsamples),
                    var.equal = TRUE)$p.value)
```

UNIFORM DISTRIBUTION

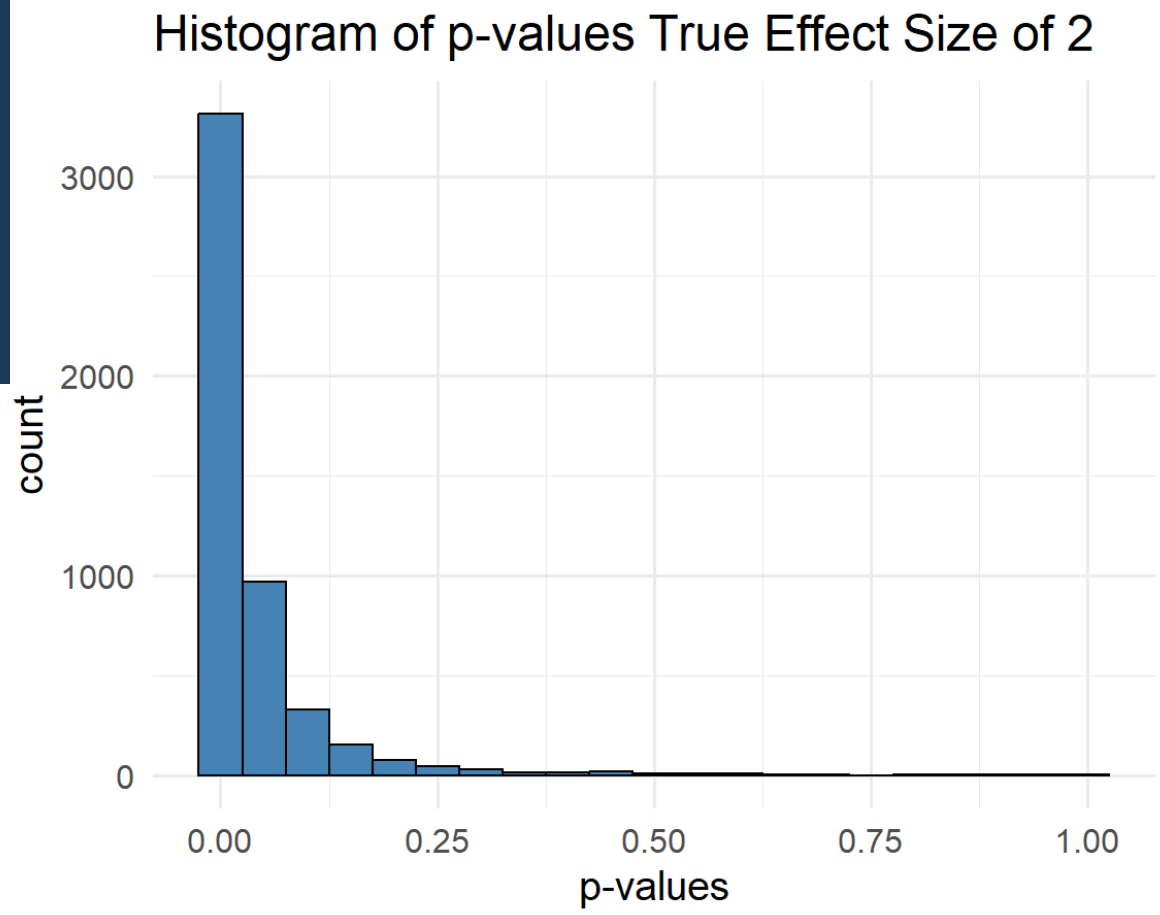


Distribution of p-values when true effect

```r
# simulated data for a 2-sample t-test
ngenes <- 5000
nsamples <- 5
effectsize <- 2

pvals <- apply(as.matrix(1:ngenes),
               1, function(a)
                 t.test(rnorm(n = nsamples),
                   rnorm(n = nsamples, mean=effectsize),
                    var.equal = TRUE)$p.value)
```

Mean of group A = 0

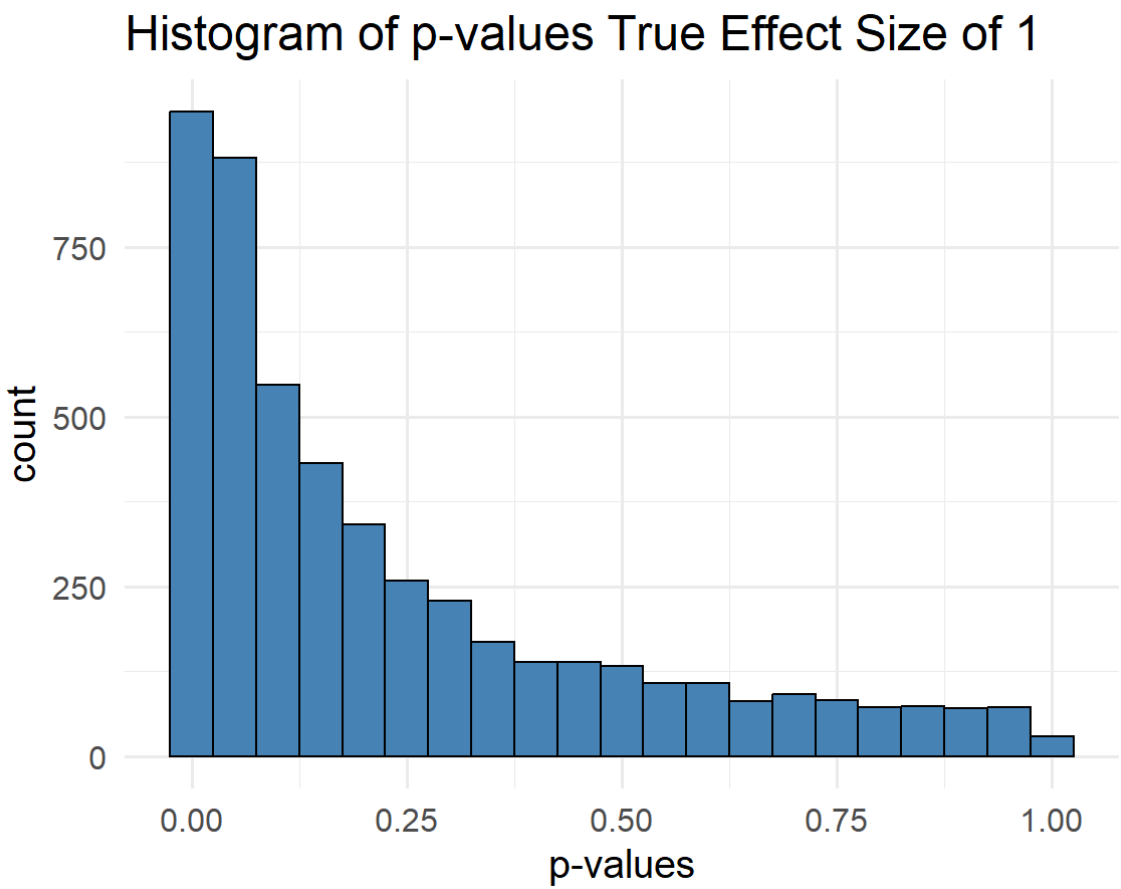
Mean of group B = 2



Distribution of p-values when true effect

```r
# simulated data for a 2-sample t-test
ngenes <- 5000
nsamples <- 5
effectsize <- 1

pvals <- apply(as.matrix(1:ngenes),
               1, function(a)
                 t.test(rnorm(n = nsamples),
                   rnorm(n = nsamples, mean=effectsize),
                    var.equal = TRUE)$p.value)
```

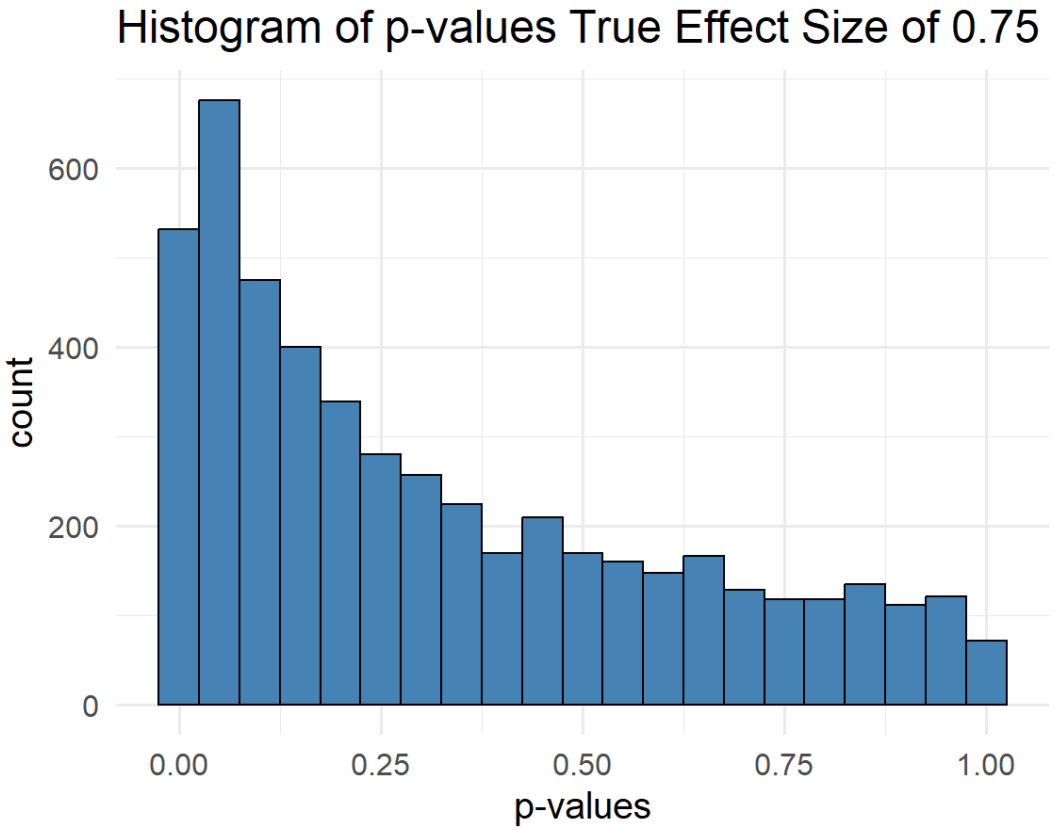
Mean of group A = 0
Mean of group B = 1



Distribution of p-values when true effect

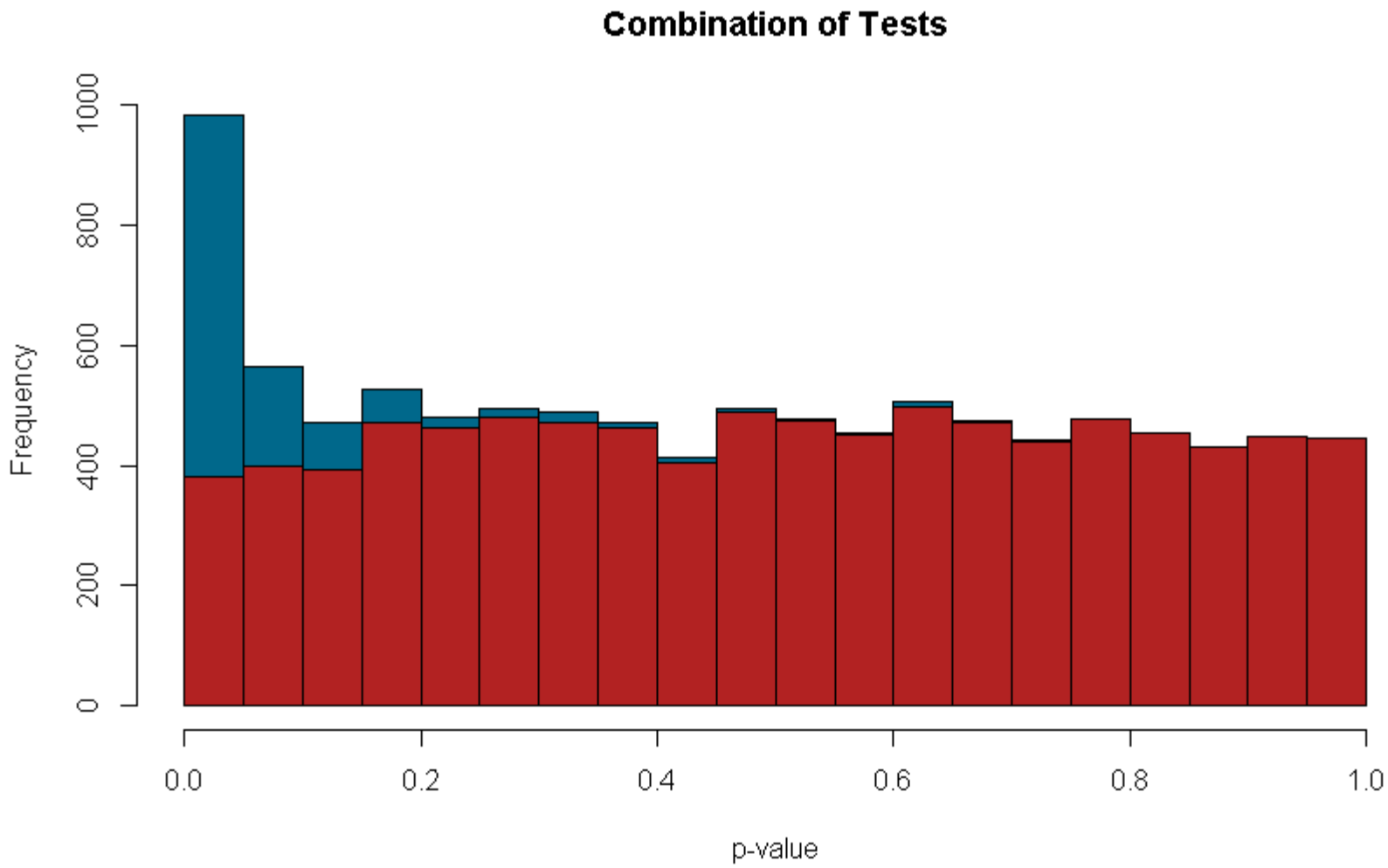
```r
# simulated data for a 2-sample t-test
ngenes <- 5000
nsamples <- 5
effectsize <- 0.75

pvals <- apply(as.matrix(1:ngenes),
               1, function(a)
                 t.test(rnorm(n = nsamples),
                   rnorm(n = nsamples, mean=effectsize),
                    var.equal = TRUE)$p.value)
```

Mean of group A = 0
Mean of group B = 0.75



Mix of results

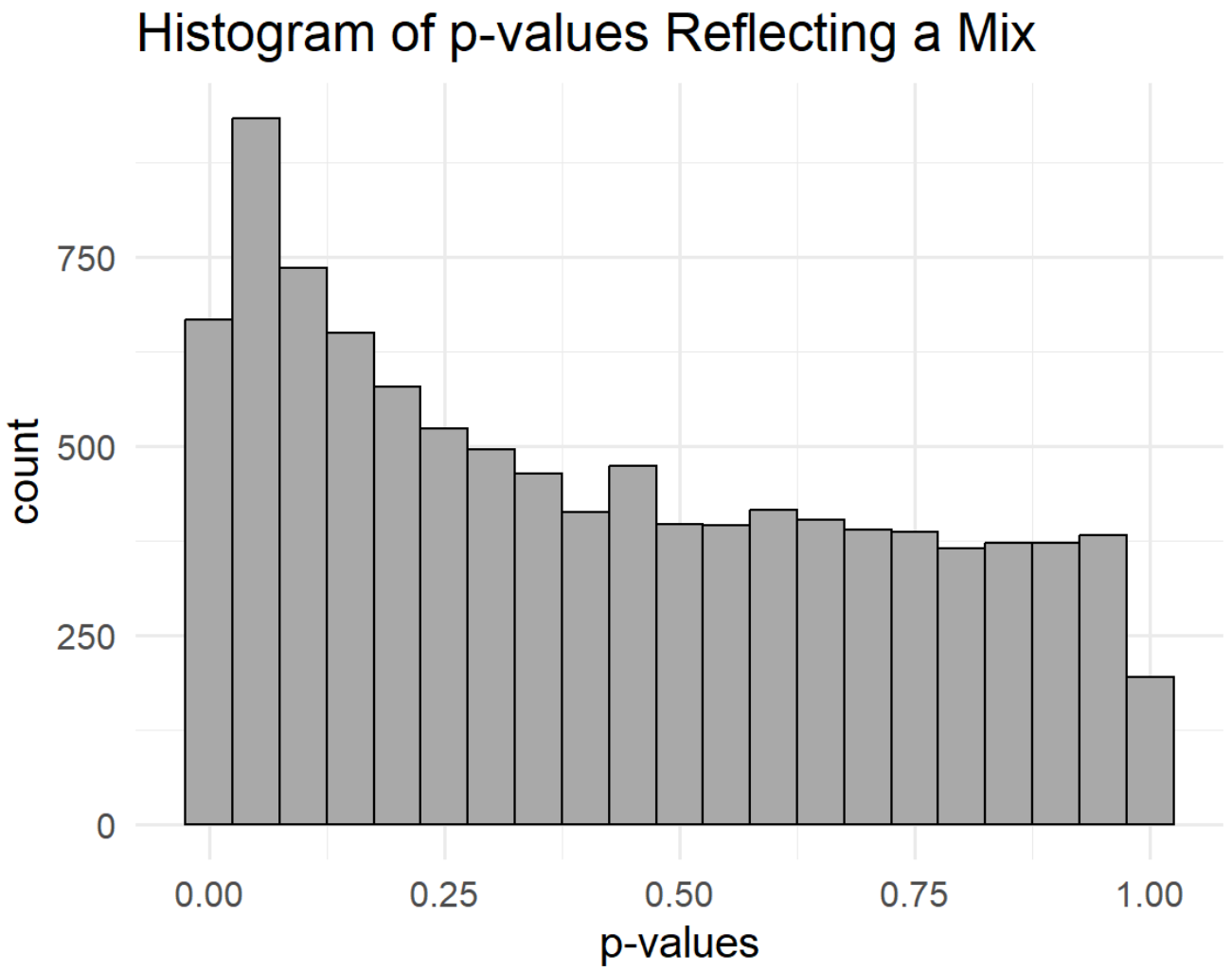
We do not have a nice color coded histogram



Mix of results

In reality, we will be looking at a distribution like this.

So how do we know what are true differences and what are not?



Q-Q Plots

The Q-Q plot is a graphical representation of the deviation of the observed p-values from the null hypothesis: the observed p-values for each SNP are sorted from largest to smallest and plotted against expected values from a uniform distribution.

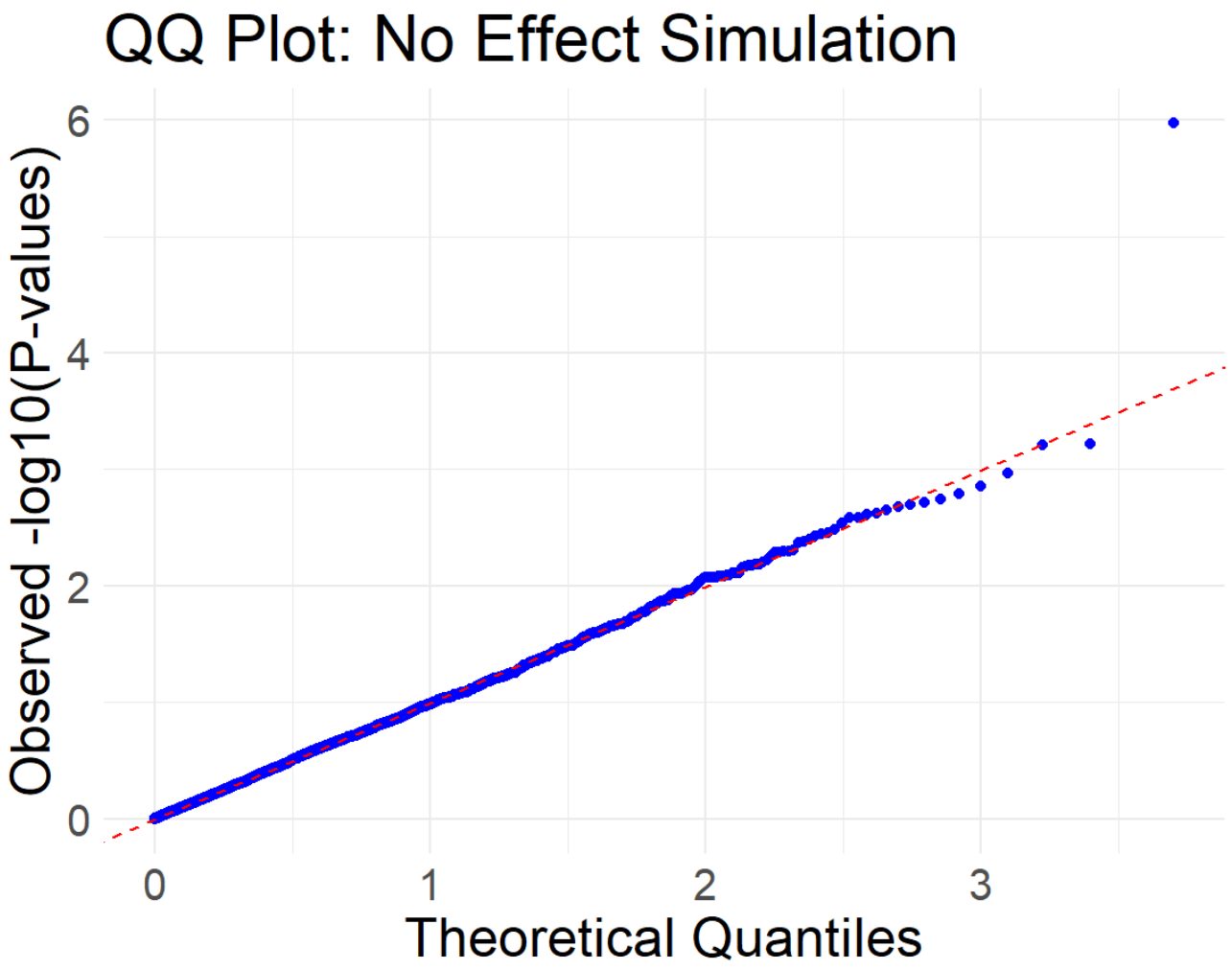
Constructing a Q-Q Plot

1. Rank observed p-values
2. Calculate expected p-value based on the value at the same rank when assuming a uniform distribution
3. Plot expected p-value (x-axis) vs. observed p-value (y-axis) – often use the negative log base 10 transformation of both
4. Examine differences between expected and observed by comparing to a line with 0 intercept and a slope of 1.

Constructing a Q-Q plot in R

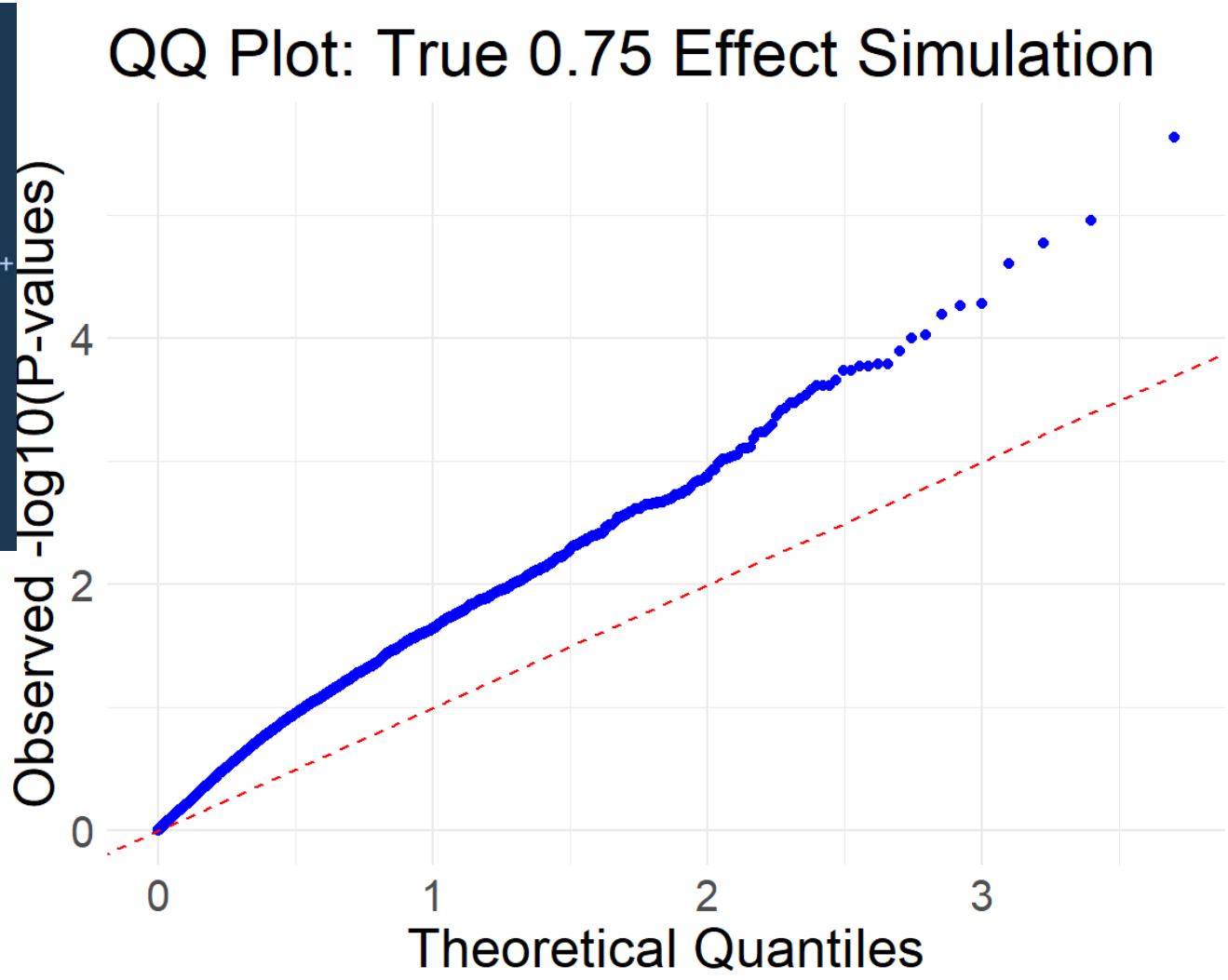
```r
# Order you p-values
pvals.null <- pvals.null[order(pvals.null)]

# Calculate theoretical quantiles
n <- length(pvals.null)
theoretical_quantiles <- -log10((1:n) / n)

# Create QQ plot
ggplot(data = NULL, aes(x = theoretical_quantiles, y =
-log10(pvals.null))) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype =
"dashed") +
  labs(x = "Theoretical Quantiles",
       y = "Observed -log10(P-values)",
       title = "QQ Plot: No Effect Simulation") +
  theme_minimal()+
   theme(text=element_text(size=20))
```

Notice the observered p-values are following the red dotted line very nicely



Constructing a Q-Q plot in R

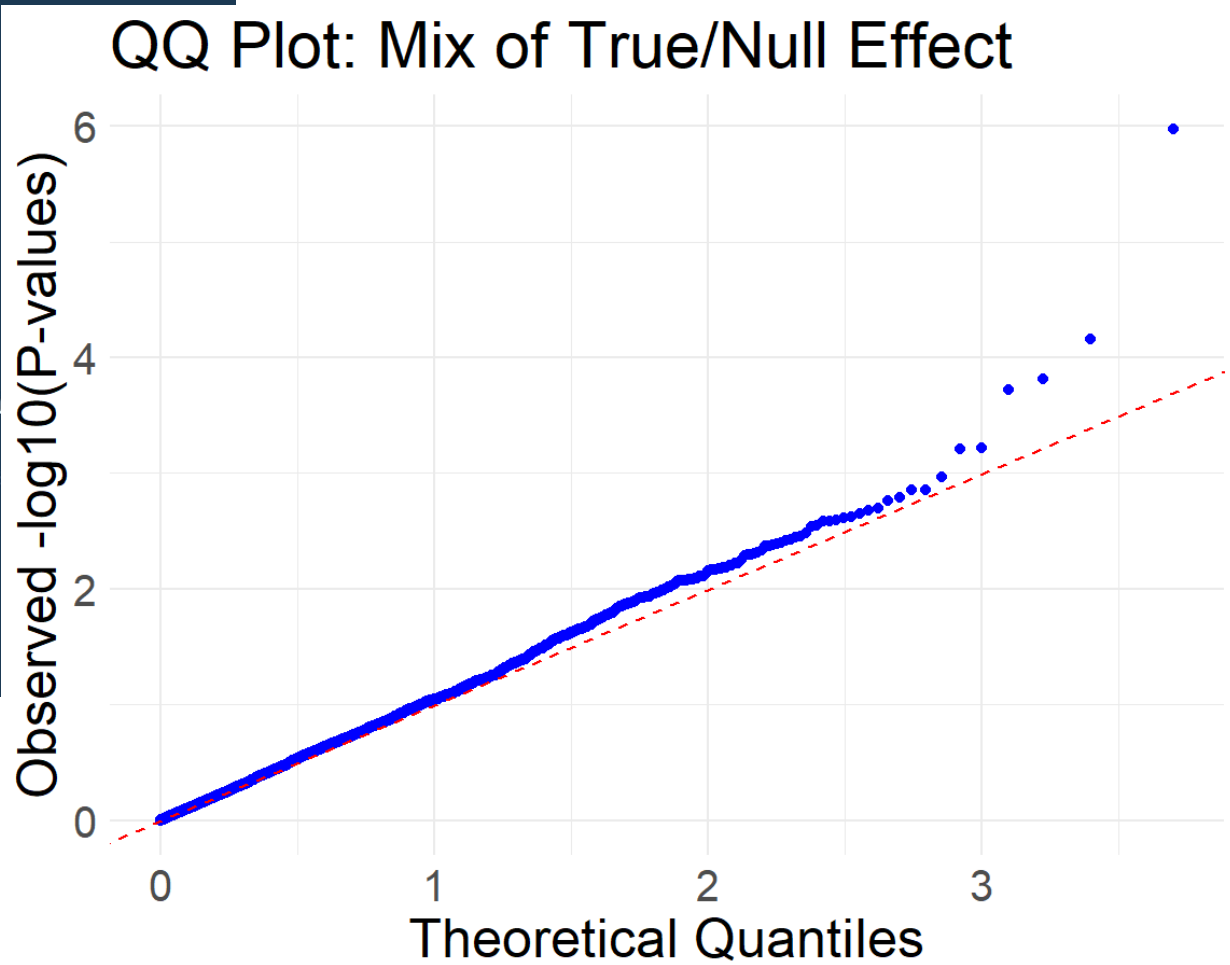
```r
# Order you p-values
pvals <- pvals[order(pvals)]

# Calculate theoretical quantiles
n <- length(pvals)
theoretical_quantiles <- -log10((1:n) / n)

# Create QQ plot
ggplot(data = NULL, aes(x = theoretical_quantiles, y = -log10(pvals))) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype =
"dashed") +
  labs(x = "Theoretical Quantiles",
       y = "Observed -log10(P-values)",
       title = "QQ Plot: True 0.75 Effect Simulation") +
  theme_minimal()+
   theme(text=element_text(size=20))
```



What people would refer to as being “inflated”

Constructing a Q-Q plot in R

```r
# get a mix of p-values where 15% come from true effect and 85% null
pvals.mix <- c(sample(x=pvals, size=0.05*ngenes),
               sample(x=pvals.null, size=0.95*ngenes))

# Order you p-values
pvals.mix <- pvals.mix[order(pvals.mix)]

# Calculate theoretical quantiles
n <- length(pvals.mix)
theoretical_quantiles <- -log10((1:n) / n)

# Create QQ plot
ggplot(data = NULL, aes(x = theoretical_quantiles, y = -log10(pva
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype =
  labs(x = "Theoretical Quantiles",
       y = "Observed -log10(P-values)",
       title = "QQ Plot: Mix of True/Null Effect") +
  theme_minimal()+
   theme(text=element_text(size=20))
```



This is what you typically should see. Closely following the line and then a spout at the tail end.

Genomic Inflation Factor (λ)

- **Assumption 1:** The majority of test statistics should follow the null distribution.
- **Assumption 2:** Markers are independent, which might not hold in terms of SNPs (high LD) or other 'omic markers.
- The **genomic inflation factor (λ):** is a metric used to assess whether a set of test statistics (e.g., from a genome-wide association study) deviates from the null hypothesis.

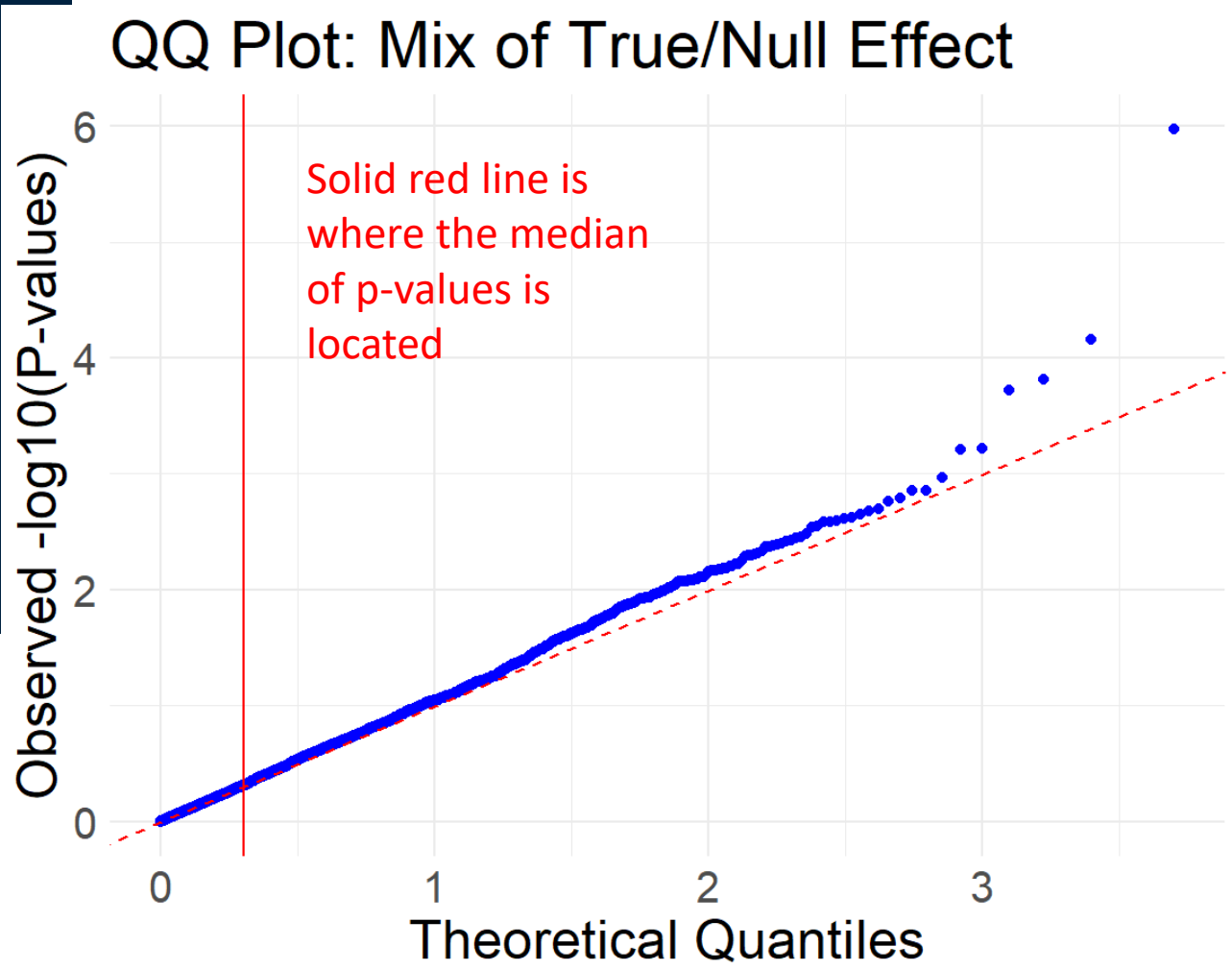
$$\lambda = \frac{Median\ of\ Observed\ \chi^2}{Median\ of\ Expected\ \chi^2\ Under\ the\ Null}$$

- Used to detect population stratification or technical artifacts that may inflate test statistics

R/Code to Calculate λ

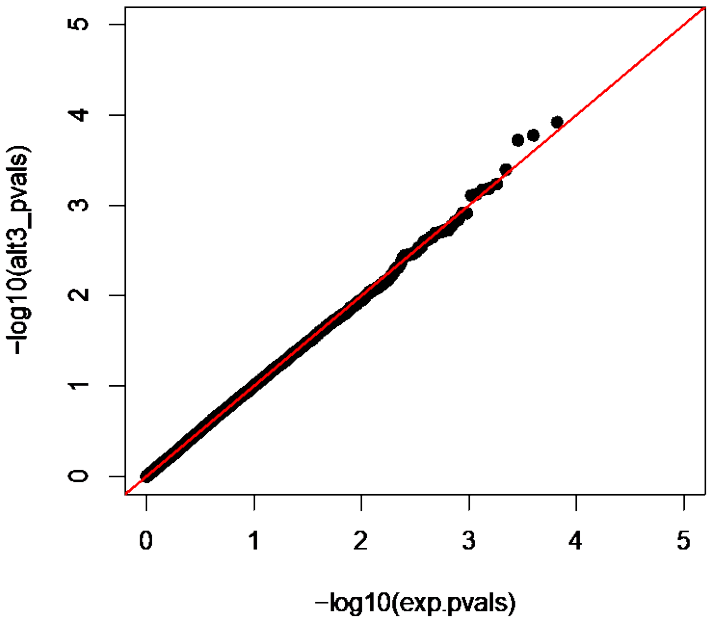
```r
# Function to calculate lambda
getLambda = function(pvals){
  chisq <- qchisq(1-pvals,1)
  #Calculate lambda gc (λgc)
  lambda = median(chisq)/qchisq(0.5,1)
  return(lambda)
}


getLambda(pvals.mix)
```

```
> getLambda(pvals.mix)
[1] 1.040191
```

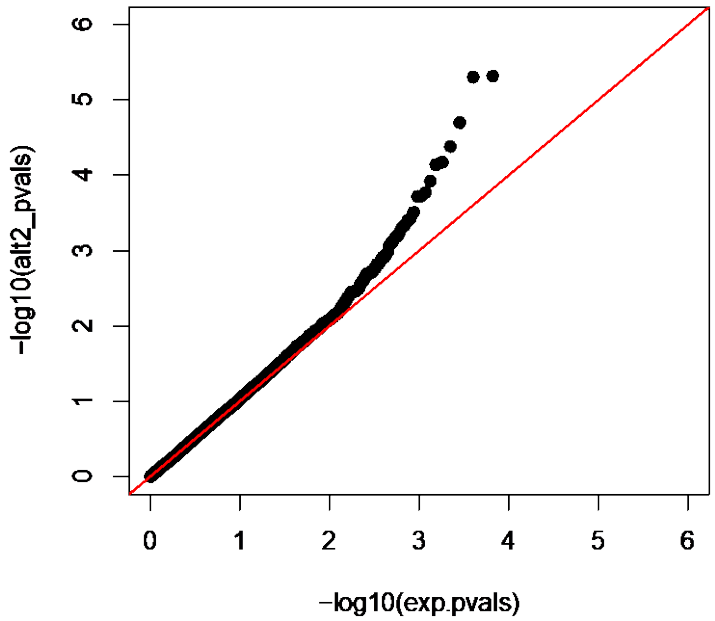


Example Q-Q plots – with Differential Expression

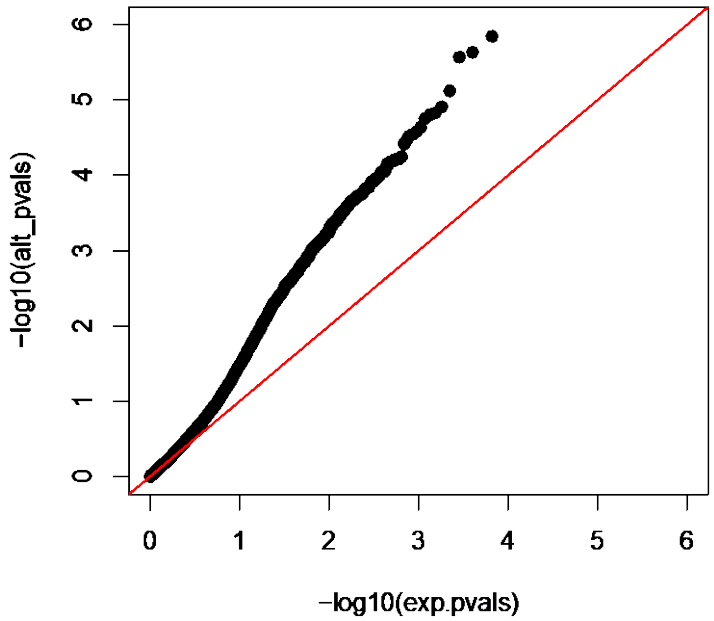
Few DE Genes



More DE Genes



Lots of DE Genes



Methods to Avoid False Positives

Multiple Testing Adjustment

Use a single ('omnibus') test to avoid a series of pairwise comparisons

- *Multiple treatments* - e.g., ANOVA, chi-square with more than two groups, multiple regression, …
- *Multiple endpoints* - multivariate statistics can handle multiple endpoints and allow omnibus testing
- *Multiple time points* - mixed linear models/repeated measures model multiple time points and allow omnibus testing

Distinction between primary and secondary (exploratory) analyses

Another way to avoid the hazards of multiple testing is to highlight one particular route through the experimentation and data analysis.

- What is the primary question being asked?
- What will be the primary endpoint that answers that question?
- What will be the primary statistical analysis of that endpoint?

The identification of a primary analysis must be finalized before the experimental data has been seen.

Secondary (exploratory) analyses

Hypothesis generating vs. hypothesis testing

- Not really answers to questions, rather they provide guidance as to what might be useful further research.

"Enjoy the result you have found by exploratory data analysis, for you will not find it again." — Stephen Senn

Look for patterns of significant results

In situations with multiple tests with no multiple testing correction:

- 2 out of 30 tests are ‘marginally’ significant = likely no true differences
- 25 out of 30 tests are significant = likely some of the tests are correct
- Do the detected differences agree with other known biology and research results?

Multiple Testing Correction

Bonferroni Correction

- Bonferroni correction is used to control the type 1 error at 5% (i.e., family-wise error rate)
- Bonferroni raises the standard of proof for all the individual tests.
- Critical p-value = 0.05/number of test
- Or, each p-value can be multiplied by the number of tests.
- BUT, this type of correction **reduces the power** and is only appropriate if each test is **independent** of the other tests

- Anderson (JASA, 2008): “[Family-wise error rate] adjustments become increasingly severe as the number of tests grows — it is inherent in controlling the probability of making a single false rejection.”

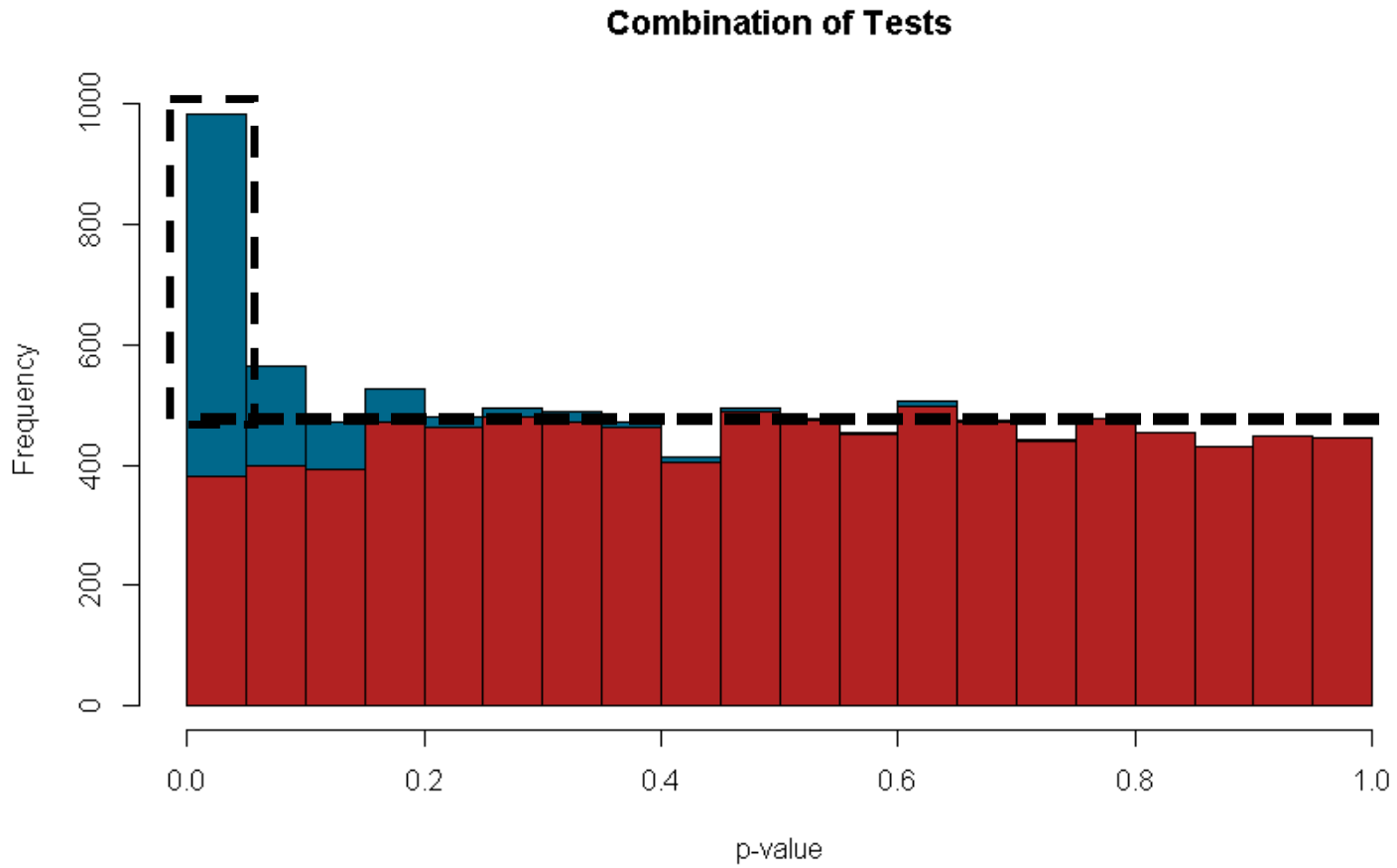
Alternative is to tolerate some small number of false positives

False Discovery Rate (FDR)

- Often for genetic studies, we use a false discovery rate (FDR) rather than a traditional p-value to help account for multiple comparisons.
- FDR is the estimated proportion of “significant” tests that are false positives at a particular threshold.
- An FDR value is calculated for each test (e.g., gene), but it is dependent on the distribution of the other test results (e.g., other genes) making a-priori power analyses difficult
- Although an FDR value is calculated for each test, it is more appropriate to report an FDR threshold rather than reporting the FDR for an individual gene.
- When we use a 10% FDR threshold for identification of ‘positive’ results, we are estimating that 10% of the ‘positive’ results are false positives.

False Discovery Rate (FDR)

- Tries to estimate your distribution of non-significant p-values (i.e. estimate that uniform background distribution)



FDR

- Benjamini–Hochberg (most common)
- Benjamini–Yekutieli
- Storey-Tibshirani

- Rcode:

```
> p.adjust(results$pvals, method="BH")
```

This gives you a vector of FDR values.

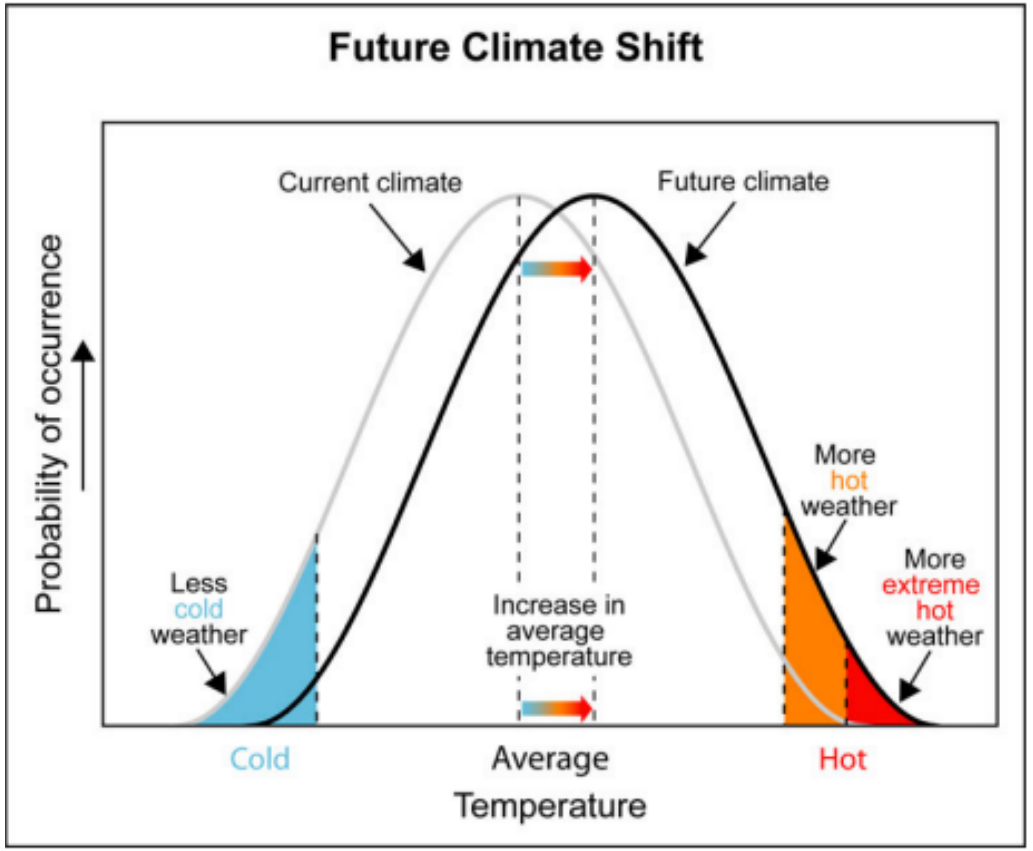
Non-Parametric Testing

Parametric vs Non-Parametric Testing

- **Parametric tests:** assume data are distributed according to a known family of probability distributions (e.g., normal).
  - If deviations from distribution of interest (e.g., Gaussian), still appropriate (robust to outliers) if sample size large (>30).
- **Non-parametric tests:** make no assumptions about the population distribution (distribution-free tests).
  - Rank-based & permutation tests
  - May be important when gross violations to distributional assumptions.

Rank-based Tests

- Use ranks of data points.
- Wilcoxon Rank Sum Test (Mann-Whitney Test) alternative for two-sample t-test.
- Evaluate if two random samples are from same distribution or if shifted in location.



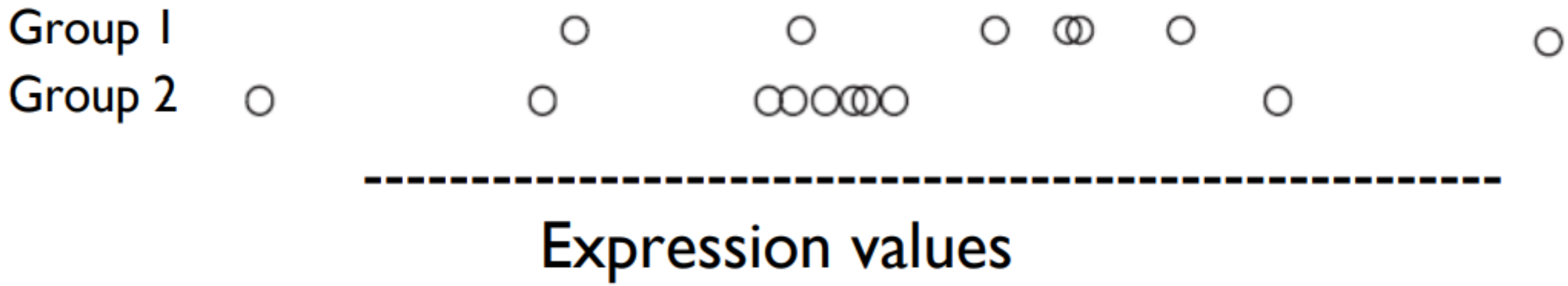
Example: Rank Test

- Suppose we have two groups and expression values (log2) for m replicate samples:

x1, x2, . . ., xm for group 1 and

y1, y2, . . ., ym for group 2

Is the distribution of the expression values for these groups significantly different (are they shifted?)



Example: Rank Test

Sample Data (n=5 per group)

	1	2	3	4	5
Group 1	-1.10	.46	-.34	.29	.82
Group 2	-1.09	.50	-1.44	1.19	.32

Combine groups and determine overall rank:

Group	1	1	1	1	1	2	2	2	2	2
Exp	-1.10	.46	-.34	.29	.82	-1.09	.50	-1.44	1.19	.32
Rank	9	4	7	6	2	8	3	10	1	5

Example: Rank Test

- Are the ranks for group 1 (or group 2) sufficiently large?
- Use test-statistic sum of the ranks for group 1
  - In our example SR* = 9 + 4 + 7 + 6 + 2 = 28
- What is the null hypothesis? Is this value extreme?
- $\binom{10}{5} = 252$ possible ranks for each group. All equally likely.

Example: Rank Test

- Calculate SR for each possible sample:

Ranks	*SR*	Probability of sample
{1,2,3,4,5}	15	1/252
{1,2,3,4,6}	16	1/252
{2,3,4,6,10}	25	1/252
{1,3,5,6,10}	25	1/252
…	…	…
{6,7,8,9,10}	40	1/252

Example: Rank Test

- Recall the definition of a p-value!

P(SR ≥ SR*) set to # (SR ≥ SR*)/252

- If the sample size is large, can use a Normal approximation

Permutation Testing

- To evaluate significance of observed test-statistic
- Evaluate all possible values of test-statistic on permuted data sets where the labels have been rearranged on the observed data
- The null hypothesis is generated from the permutations (do not need to assume distribution)

Example: Permutation Test

- Calculate t-statistic on previous example

Group	1	1	1	1	1	2	2	2	2	2
Exp	-1.10	.46	-.34	.29	.82	-1.09	.50	-1.44	1.19	.32
Rank	9	4	7	6	2	8	3	10	1	5
Mean	.026					-0.104				

- Two-sample t-statistic $t^* = 0.22$
- Do not assume t-distribution, use permutations

Example: Permutation Test

- Permuted data set 1

Group	1	2	1	2	2	2	1	1	2	1
Exp	-1.10	.46	-.34	.29	.82	-1.09	.50	-1.44	1.19	.32
Rank	9	4	7	6	2	8	3	10	1	5

Two-sample t-statistic $t_1 = -1.37$

- Permuted data set 2

Group	2	2	2	1	1	1	2	2	1	1
Exp	-1.10	.46	-.34	.29	.82	-1.09	.50	-1.44	1.19	.32
Rank	9	4	7	6	2	8	3	10	1	5

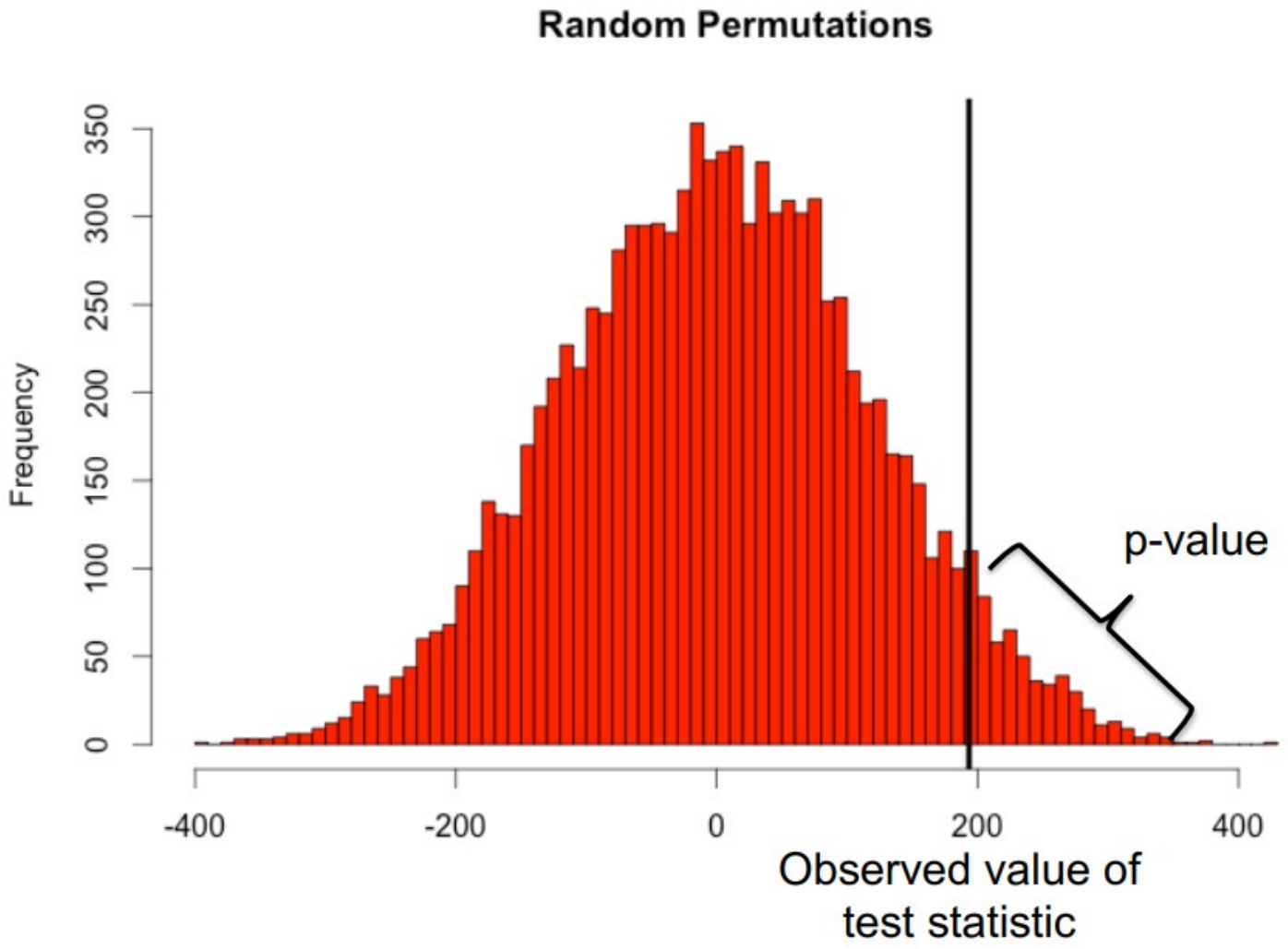
Two-sample t-statistic $t_2 = 2.25$

- …. repeat many times…….

Example: Permutation Test

- If sample small enough, all permutations (p) can be evaluated.
- Otherwise, sample randomly (e.g., 10000 times) from all possible permutations.
- Recall the definition of a p-value!
- $P(t^* \geq t) = \#\ (t_p \geq t\ ) / \#$ of permutations
- Possible p-values for both examples are discrete
- For example: if your t* is lower than all permuted t-values on 10,000 permutations, you would report a p-value < 1/10,000 (or < 1e-04)

Permutation p-value



Issues

Exchangeability

- The joint distributions is invariant to permutations:

  Example: distribution of $(X_1, X_2, X_3, X_4, X_5)$ is exchangeable with distribution of $(X_3, X_4, X_1, X_5, X_2)$

- Works if data are independent and identically distributed (IID)
- May fail when there are sub-groups in the data, repeated measurements, dependencies between observations

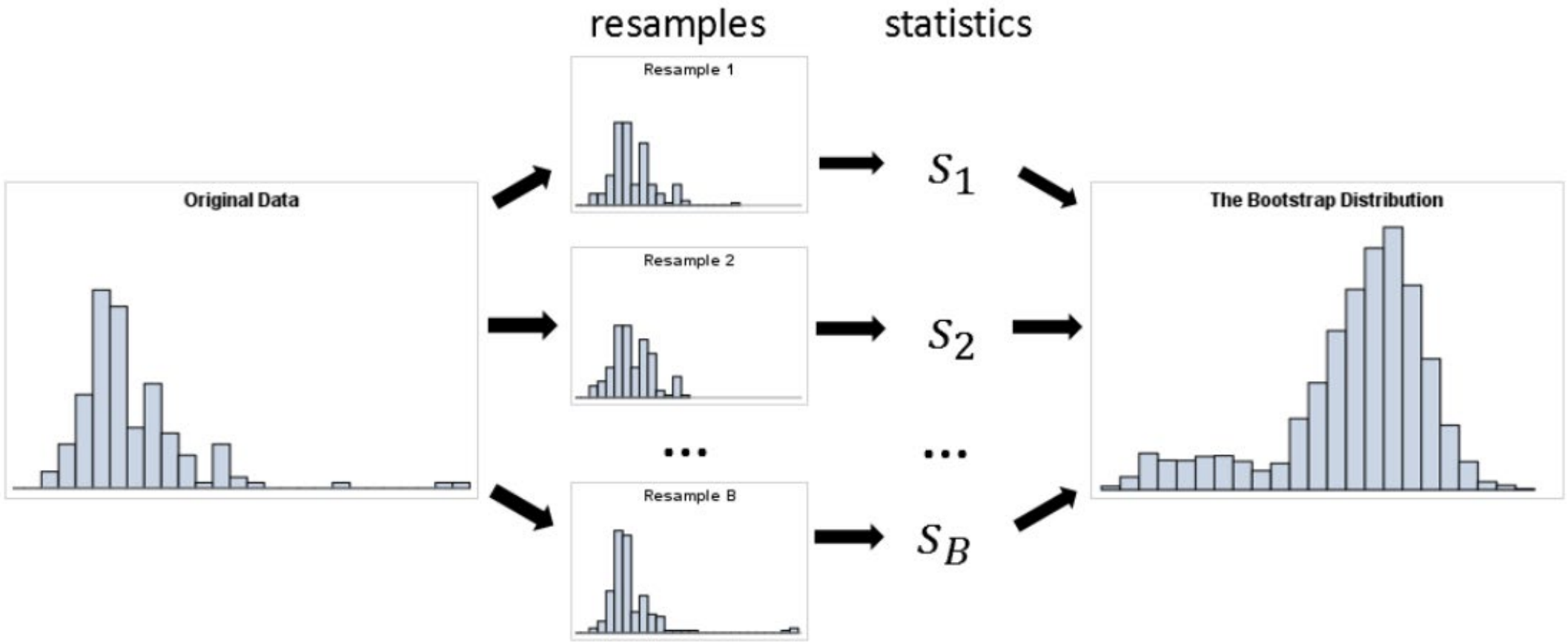
Issues

- Exchangeability (cont.)
  - e.g., observe heads/tails, but there are two different coins
  - Need to stratify permutations by sub-group
  - Perform multi-block permutations
- Practical Concerns
  - n too small – not enough permutations (e.g., n=10, 2 groups – 252 permutations)
  - n big – cannot do all, randomly sample

Bootstrapping

- What if you want to get a confidence interval or standard error on an estimate that is not parametric?
- Draw sample
  - *with* replacement – original sample size n
  - Repeat to obtain B bootstrap samples – B = 100, 1000
  - For each b=1,2, … B bootstrap samples, calculate statistic S_b
- Obtain sampling distribution of statistic S – Estimate standard error, confidence interval, etc.

Example: Bootstrapping



Bootstrap Properties

- Sample variance of statistic S converges to true variance as B->infinity

- Straightforward
  - Check stability of results
  - Useful for complex estimators

Permutation/Bootstrap Note

Save seed to reproduce results!

Conclusions

What did we learn?

- Multiple testing burden needs to be dealt with
- Histogram of p-values and Q-Q plots are good ways to visualize p-values
- There are various ways to correct for multiple testing (Bonferroni, FDR)
- Permutations and bootstrapping is a great way to tackle non-parametric tests