PQHS 452
Multiple Testing and Statistical Power

The Lady Tasting Tea

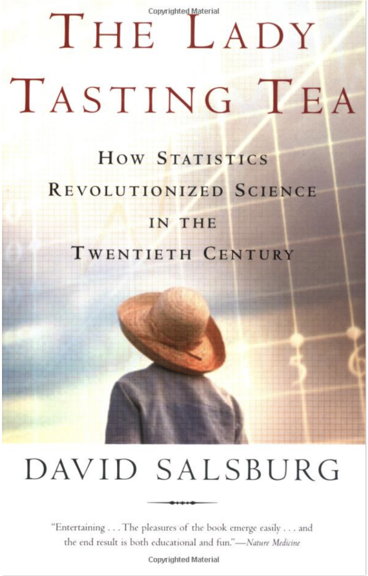
- It was a summer afternoon in Cambridge, England, in the 1920s.
- A groups of university dons, their wifes, and some guests were having afternoon tea.
- A lady was insisting that tea tasted different depending upon whether *the tea was poured into the milk* OR *the milk was poured into the tea*.

The Lady Tasting Tea

Fisher in 1913

- “Sheer nonsense”, the scientific minds among the men scoffed at this.
- A thin, short man, with thick glasses, Ronald Fisher, pounced on the problem: “Let us test the proposition!”

ASA Statement on p-values



Copyrighted Material

THE LADY TASTING TEA

HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY

DAVID SALSBURG

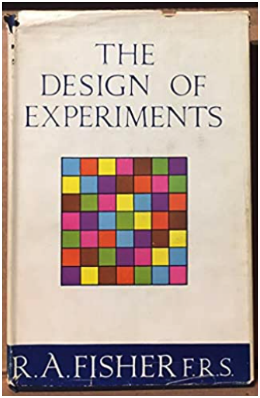
"Entertaining . . . The pleasures of the book emerge easily . . . and the end result is both educational and fun."—*Nature Medicine*

Copyrighted Material

Hypothesis Testing

- Fisher's notion of a *null hypothesis*
  — Null hypothesis
  — Popularize p-value
- Neyman-Pearson Lemma
  — Error of the 2nd kind
  — Alternative/competing hypothesis
  — Power function

Most influential books on statistical methods



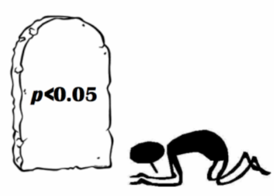
- **Statistical Methods for Research Workers**
- **The Design of Experiments**

“...the best thing about being a statistician...”

John Wilder Tukey

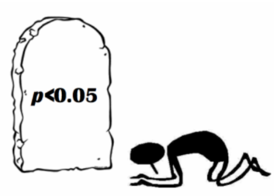
“... is that you get to play in everyone’s backyard.”

Misuse of p-value



- Q: Why do so many colleges and grad schools teach p = 0.05?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
- A: Because that's what they were taught in college or grad school.

Misuse of p-value



- Q: Why do so many colleges and grad schools teach p = 0.05?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
- A: Because that's what they were taught in college or grad school.

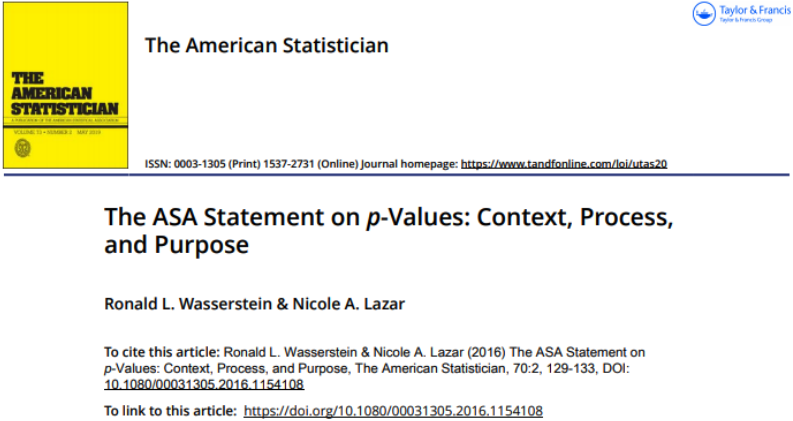
"We teach it because it's what we do; we do it because it's what we teach."

Fisher's words in SMRW

STATISTICAL METHODS FOR RESEARCH WORKERS FISHER FOURTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER FIFTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER SIXTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER SIXTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER SEVENTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER EIGHTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER NINTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER TENTH EDITION OLIVER AND BOYD
STATISTICAL METHODS FOR RESEARCH WORKERS FISHER ELEVENTH EDITION OLIVER AND BOYD

"Personally, the writer prefers to set a low standard of significance at 5 percentage point. . . A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

ASA Statement on p-values



THE AMERICAN STATISTICIAN

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: https://doi.org/10.1080/00031305.2016.1154108

pop quiz

Which(s) of the following statements is/are reasonable?

- p-value is a probability.
- $p > 0.05$ is the probability that the null hypothesis is true.
- 1 minus the p-value is the probability that the alternative hypothesis is true.
- A statistically significant test result ($p \leq 0.05$) means that the test hypothesis is false or should be rejected.
- A p-value greater than 0.05 means that no effect was observed.

The status quo

Informally, a p-value is the probability **under a specified statistical model** that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be *equal to or more extreme* than its observed value.

Six principles of p-value

- 1. P-values can indicate how incompatible the data are with a specified statistical model.
  — The most common context is a model (under a set of assumptions): H_0
  — Often H_0 postulates the absence of an effect (e.g. no difference between two groups)
  — The smaller the p-value, the greater the incompatibility of the data with H_0
  — Incompatibility casting doubt on H_0

Six principles of p-value

- 1. P-values can indicate how incompatible the data are with a specified statistical model.
  — The most common context is a model (under a set of assumptions): H_0
  — Often H_0 postulates the absence of an effect (e.g. no difference between two groups)
  — The smaller the p-value, the greater the incompatibility of the data with H_0
  — Incompatibility casting doubt on H_0
- 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
  — Never turn a p-value into a statement about the truth of H_0
  — p-value is a statement about the **relationship** between the data and H_0, NOT about the **explanation** (H_0) itself.

Six principles of p-value (cont'd)

- 3. Scientific conclusions and business or policy decisions should NOT be based only on whether a p-value passes a specific threshold.
  — "bright-line" rule (e.g. $p < 0.05$ alone) can lead to erroneous beliefs and poor decision making.
  — A conclusion does not immediately become "true" on one side of the divide and "false" on the other.
  — Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.
  — Using $p < 0.05$ alone as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.

Six principles of p-value (cont'd)

- 3. Scientific conclusions and business or policy decisions should NOT be based only on whether a p-value passes a specific threshold.
  — "bright-line" rule (e.g. $p < 0.05$ alone) can lead to erroneous beliefs and poor decision making.
  — A conclusion does not immediately become "true" on one side of the divide and "false" on the other.
  — Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.
  — Using $p < 0.05$ alone as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.
- 4. Proper inference requires full reporting and transparency
  — number of hypotheses explored, all data collection decisions, all statistical analyses conducted
  — No "cherry-picking"

Six principles of p-value (cont'd)

- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  — $pval \neq$effect size
  — Statistical sig. vs. biological sig.

Six principles of p-value (cont'd)

- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  — $pval \neq$effect size
  — Statistical sig. vs. biological sig.
- 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Usage of p-value

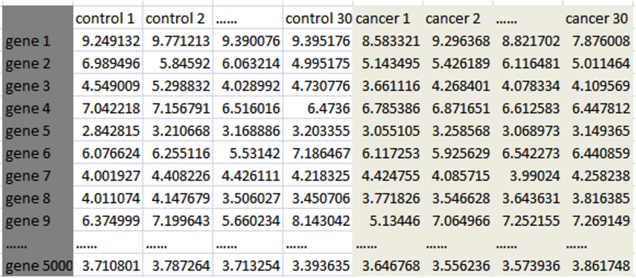
- **Good statistical practice** is an integral part of **good scientific practice**.
  — study design and conduct, summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting, proper logical understanding of results.

Usage of p-value

- **Good statistical practice** is an integral part of **good scientific practice**.
  — study design and conduct, summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting, proper logical understanding of results.
- **No single index should substitute for scientific reasoning.**

Hypothesis testing in genomics

Gene/protein/metabolite expression data.



	control 1	control 2	......	control 30	cancer 1	cancer 2	......	cancer 30
gene 1	9.249132	9.771213	9.390076	9.395176	8.583321	9.296368	8.821702	7.876008
gene 2	6.989496	5.84592	6.063214	4.995175	5.143495	5.426189	6.116481	5.011464
gene 3	4.549009	5.298832	4.028992	4.730776	3.661116	4.268401	4.078334	4.109569
gene 4	7.042218	7.156791	6.516016	6.4736	6.785386	6.871651	6.612583	6.447812
gene 5	2.842815	3.210668	3.168886	3.203355	3.055105	3.258568	3.068973	3.149365
gene 6	6.076624	6.255116	5.53142	7.186467	6.117253	5.925629	6.542273	6.440859
gene 7	4.001927	4.408226	4.426111	4.218325	4.424755	4.085715	3.99024	4.258238
gene 8	4.011074	4.147679	3.506027	3.450706	3.771826	3.546628	3.643631	3.816385
gene 9	6.374999	7.199643	5.660234	8.143042	5.13446	7.064966	7.252155	7.269149
......	......	......	......	......	......	......	......	......
gene 5000	3.710801	3.787264	3.713254	3.393635	3.646768	3.556236	3.573936	3.861748

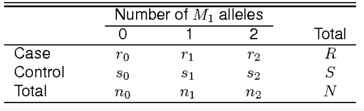
After all the pre-processing, we have a feature by sample matrix of expression indices.

It is like an molecular “fingerprint” of each sample.

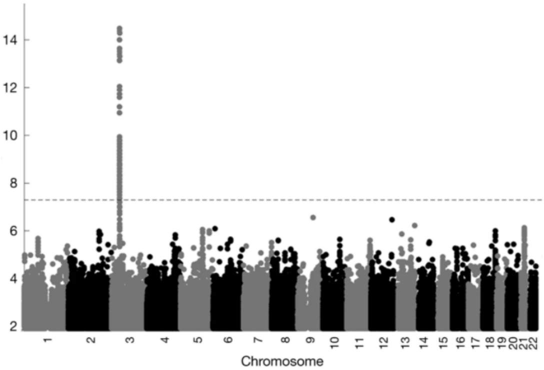
The most common use: to find biomarkers of a disease.

Hypothesis testing in genomics

Genetics/SNP data.



	Number of M_1 alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N



The problem: multiple testing

How does the problem of multiple testing arise?

Let us use T to denote the random variable (e.g. test statistics), use $F(t)$ to denote its cumulative distribution function (CDF). By definition, we have $F(t) \equiv Pr(T < t)$ for all t.

$F()$ is invertible (in general), we can derive the distribution of the random p-value $P = F(T)$ (or symmetrically $1 - F(T)$) as follows:

$$Pr(P < p) = Pr(F(T) < p) = Pr(T < F^{-1}(p)) = F(F^{-1}(p)) = p$$

Now we can conclude that the distribution of p-value as a RV P is uniform on $[0, 1]$.

The problem: multiple testing

Theorem

Under the null hypothesis, p-values distribute uniformly on $[0, 1]$.

Suppose in a GWAS studies with 100,000 SNPs are tested for genetic association separately, you found 6,000 significant ($p < 0.05$) loci.
Is that good?

The problem: multiple testing

Theorem

Under the null hypothesis, p-values distribute uniformly on $[0, 1]$.

Suppose in a GWAS studies with 100,000 SNPs are tested for genetic association separately, you found 6,000 significant ($p < 0.05$) loci.
Is that good?
NO! Because even if there is no genetic association at all (H_0 holds), you'll observe $100,000 \times 0.05 = 5,000$ significant loci.
So... out of the 6,000 significant loci you identified, 5,000 could be false positives.

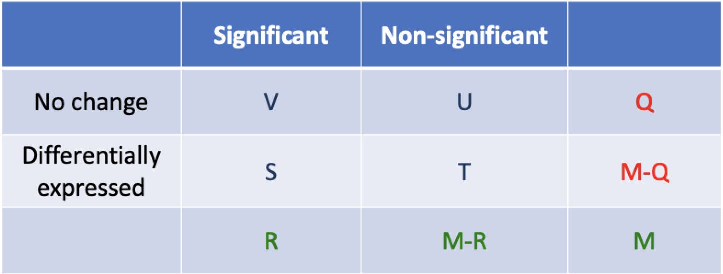
The problem: multiple testing

Theorem

Under the null hypothesis, p-values distribute uniformly on $[0, 1]$.

Suppose in a GWAS studies with 100,000 SNPs are tested for genetic association separately, you found 6,000 significant ($p < 0.05$) loci.
Is that good?
NO! Because even if there is no genetic association at all (H_0 holds), you'll observe $100,000 \times 0.05 = 5,000$ significant loci.
So... out of the 6,000 significant loci you identified, 5,000 could be false positives.
We use **False Discovery Rate** (FDR) to conceptualize the rate of type I errors. Here, $FDR = \frac{5000}{6000} = 0.83$ is indeed miserable.

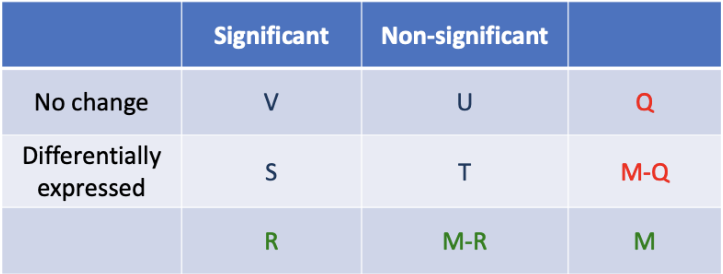
General considerations



	Significant	Non-significant	
No change	V	U	Q
Differentially expressed	S	T	M-Q
	R	M-R	M

Simultaneously test M hypotheses.
Q is # true null – genes that didn't change (unobserved)
R is # rejected – genes called significant (observed)
U, V, T, S are unobservable random variables.
V: number of type-I errors; T: number of type-II errors.

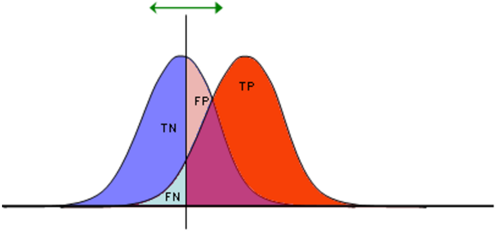
General considerations

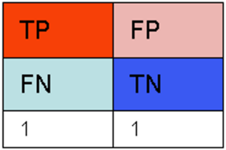


	Significant	Non-significant	
No change	V	U	Q
Differentially expressed	S	T	M-Q
	R	M-R	M

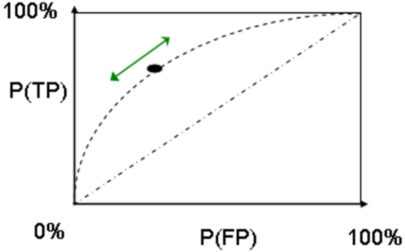
Sensitivity: E[S/(M-Q)]
Specificity: E[U/Q]
False discovery rate (FDR) = E(V/R)

General considerations

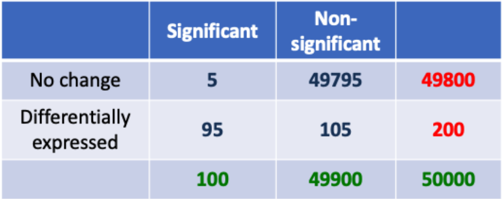




TP	FP
FN	TN
1	1

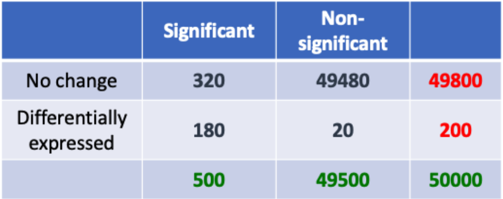


General considerations



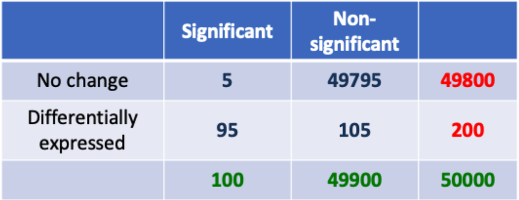
	Significant	Non-significant	
No change	5	49795	49800
Differentially expressed	95	105	200
	100	49900	50000

It makes more sense than this, which leans too heavily towards sensitivity:



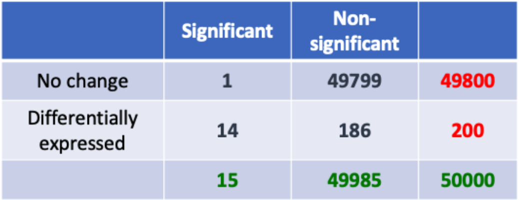
	Significant	Non-significant	
No change	320	49480	49800
Differentially expressed	180	20	200
	500	49500	50000

General considerations



	Significant	Non-significant	
No change	5	49795	49800
Differentially expressed	95	105	200
	100	49900	50000

It makes more sense than this, which leans too heavily towards specificity:



	Significant	Non-significant	
No change	1	49799	49800
Differentially expressed	14	186	200
	15	49985	50000

Family-wise error rate (FWER)

When we have multiple tests, let G be the number of true nulls called significant (false positives). Then,

$$FWER = Pr(G \geq 1) = 1 - Pr(G = 0)$$

“Family”: a group of hypothesis that are similar in prupose, and need to be jointly accurate.

Bonferroni correction is one version of FWER control.

Bonferroni correction

Suppose we have m tests, $m = 1, 2, ..., M$.
Bonferroni correction: An easy and popular approach to adjust the significance level of each test so as to preserve the FWER:

$$\begin{aligned}\alpha &= P(\text{reject at least one } H_0^{(m)} | H_0^{(m)} \text{ is true for all } m) \\ &= P(\cup_m \{\text{reject } H_0^{(m)} | H_0^{(m)} \text{ is true}\}) \\ &\leq \sum_m P(\text{reject } H_0^{(m)} | H_0^{(m)} \text{ is true}) \\ &= M\alpha'\end{aligned}$$

FWER can be kept $< \alpha$, if each individual test has significance level α / M.
e.g. $\alpha = 0.01$, and $M = 500,000$, then $\alpha' = 2 \times 10^{-8}$.
Bonferroni correction is the simplest and most conservative approach.

Other methods in multiple testing

- FDR - (Benjamini and Hochberg) BH procedure
- q-value, pFDR
- Efron's Local FDR

Back on the two types of errors

- **Type I Error**: False Positive. Reject H_0 when there is in fact NO true difference.
- **Type II Error**: False Negative. Not reject the null hypothesis when there IS in fact true difference.

Statistical Power

- Statistical power is the probability that the test correctly rejects the null hypothesis.

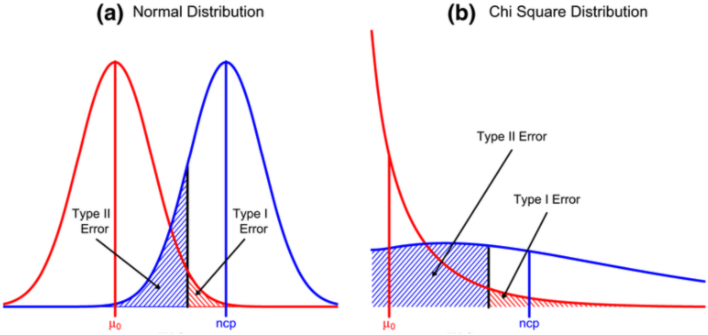
In other words: Given the alternative hypothesis (H_A) is the underlying truth, the probability that we'll reject H_0 is called statistical power.

Power $=$ 1 - Type II error.

Other puzzle pieces needed for power evaluation

- Significance level (α)
- Sample size
- Effect size
- Variability

Primary components of power



Power $= 1 -$ shaded blue area.

Power calculation example 1

z-test

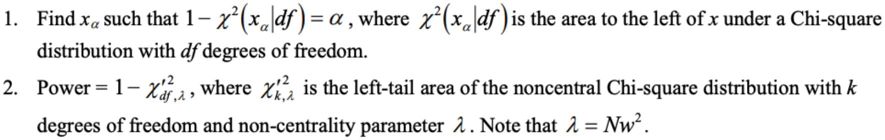
Denote $P(Z \leq z) = \Phi(z)$, the area to the left of z under the standard Normal curve. Define effect size $\triangle = \delta = \frac{\mu - \mu_0}{\sigma}$. Consider $\bar{x} \sim N(\mu, \sigma_{\bar{x}}^2)$:

$$\begin{aligned}
\text{Power} &= P_\mu(\bar{x} > \mu_0 + z_{1-\alpha/2}\sigma_{\bar{x}}) + P_\mu(\bar{x} < \mu_0 - z_{1-\alpha/2}\sigma_{\bar{x}}) \\
&= P(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} > \frac{\mu_0 - \mu}{\sigma_{\bar{x}}} + z_{1-\alpha/2}) + P(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{\mu_0 - \mu}{\sigma_{\bar{x}}} - z_{1-\alpha/2}) \\
&= \Phi(\sqrt{n}\Delta - z_{1-\alpha/2}) + \Phi(-\sqrt{n}\Delta - z_{1-\alpha/2})
\end{aligned}$$

The 2nd part is often ignored due to extremely small resulting value.

Power calculation example 2

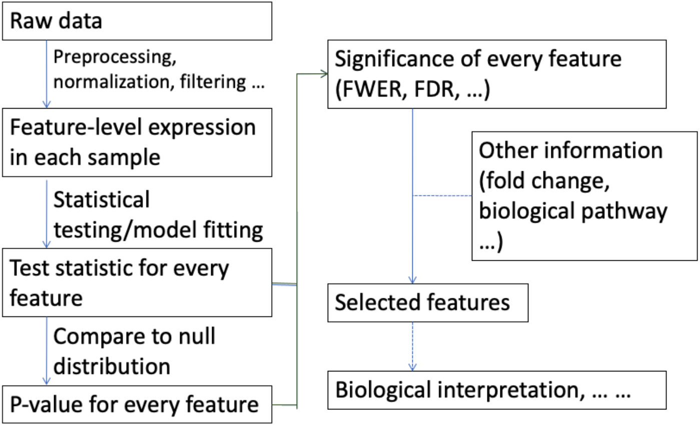
Chi-square test



1. Find x_α such that $1 - \chi^2\left(x_\alpha \middle| df\right) = \alpha$, where $\chi^2\left(x_\alpha \middle| df\right)$ is the area to the left of x under a Chi-square distribution with *df* degrees of freedom.
2. Power $= 1 - \chi'^2_{df,\lambda}$, where $\chi'^2_{k,\lambda}$ is the left-tail area of the noncentral Chi-square distribution with k degrees of freedom and non-centrality parameter λ. Note that $\lambda = Nw^2$.

where N is the total count in all the cells. w is the effect size.

Overview of genomics data analysis workflow



Raw data
Preprocessing, normalization, filtering …
Feature-level expression in each sample
Statistical testing/model fitting
Test statistic for every feature
Compare to null distribution
P-value for every feature
Significance of every feature (FWER, FDR, …)
Other information (fold change, biological pathway …)
Selected features
Biological interpretation, … …