Common Errors in Statistics: Medicine / Bioinformatics

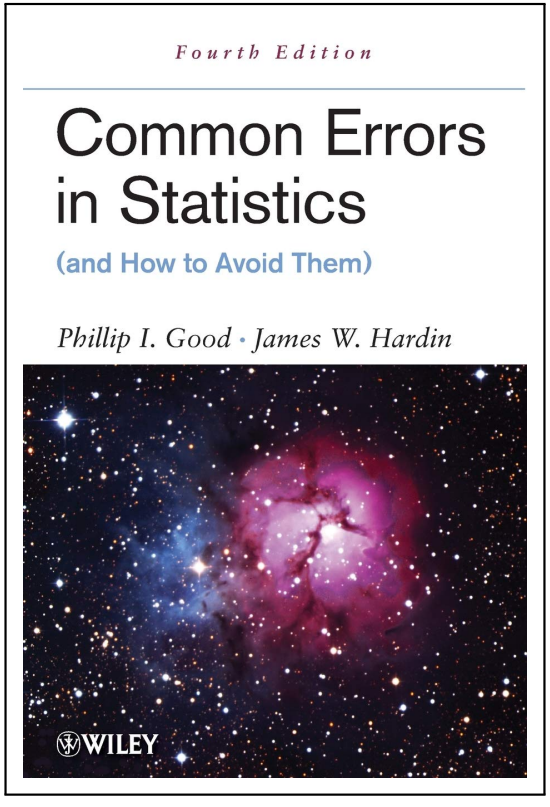
Mitch Murphy

alphataubio@gmail.com

BCH 8109

2025/10/17

Book Recommendation: Common Errors in Statistics (And How to Avoid Them)



PART I FOUNDATIONS

1. Sources of Error
2. Hypotheses: The Why of Your Research
3. Collecting Data

PART II STATISTICAL ANALYSIS

4. Data Quality Assessment
5. Estimation
6. Testing Hypotheses: Choosing a Test Statistic
7. Strengths and Limitations of Statistical Procedures
8. Reporting Your Results
9. Interpreting Reports
10. Graphics

PART III BUILDING A MODEL

11. Univariate Regression
12. Alternate Methods of Regression
13. Multivariable Regression
14. Modelling Counts and Correlated Data
15. Validation

Free PDF download from uOttawa OMNI:

https://onlinelibrary-wiley-com.proxy.bib.uottawa.ca/doi/book/10.1002/9781118360125

nature communications

Article

https://doi.org/10.1038/s41467-023-36150-7

Chromatin complex dependencies reveal targeting opportunities in leukemia

Received: 19 September 2022

Accepted: 18 January 2023

Published online: 27 January 2023

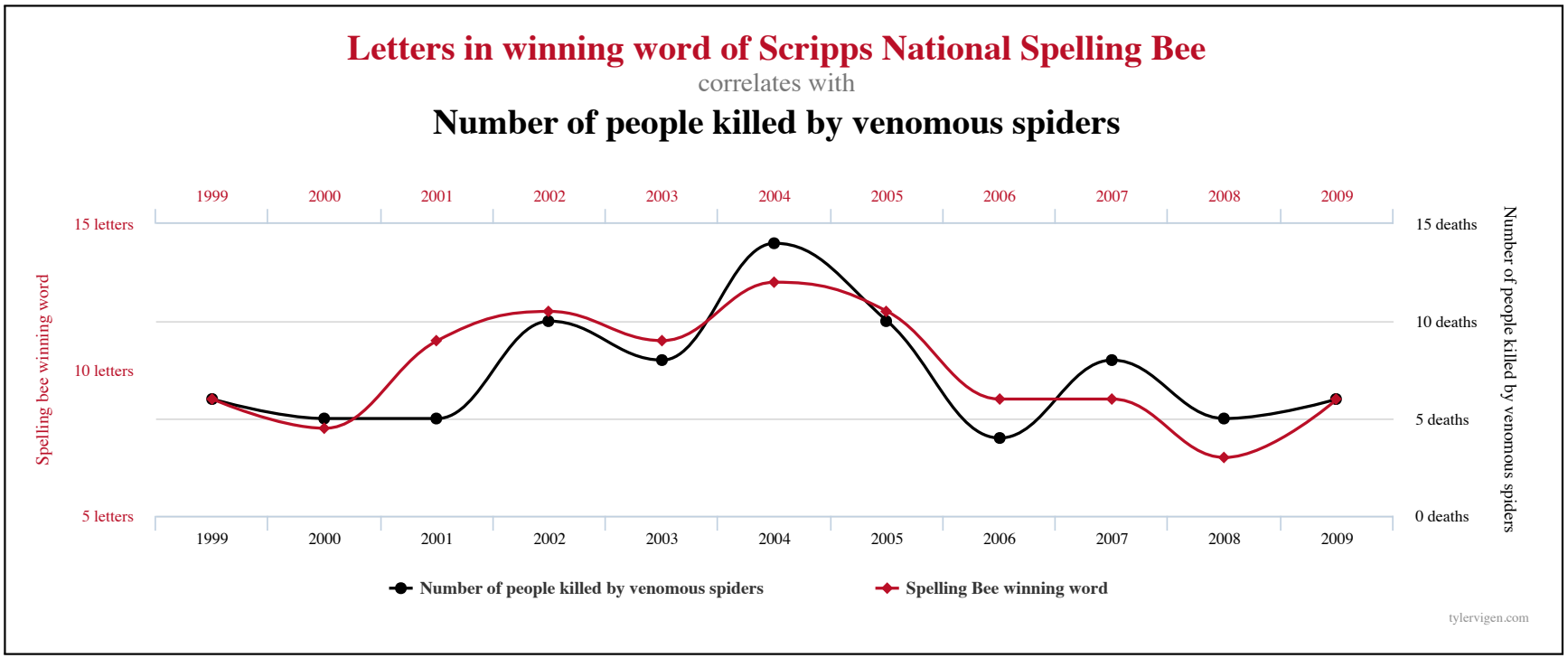
Check for updates

Fadi J. Najm[1], Peter DeWeirdt[2], Molly M. Moore[1], Samantha M. Bevill[1,3,4], Chadi A. El Farran[1,3,4], Kevin A. Macias[1,3,4], Mudra Hegde[2], Amanda L. Waterbury[1,5], Brian B. Liau[1,5], Peter van Galen[6], John G. Doench[2] & Bradley E. Bernstein[1,3,4]

Chromatin regulators are frequently mutated in human cancer and are attractive drug targets. They include diverse proteins that share functional domains and assemble into related multi-subunit complexes. To investigate functional relationships among these regulators, here we apply combinatorial CRISPR knockouts (KOs) to test over 35,000 gene-gene pairings in leukemia cells, using a library of over 300,000 constructs. Top pairs that demonstrate either compensatory non-lethal interactions or synergistic lethality enrich for paralogs and targets that occupy the same protein complex. The screen highlights protein complex dependencies not apparent in single KO screens, for example MCM histone exchange, the nucleosome remodeling and deacetylase (NuRD) complex, and HBO1 (KAT7) complex. We explore two approaches to NuRD complex inactivation. Paralog and non-paralog combinations of the KAT7 complex emerge as synergistic lethal and specifically nominate the ING5 PHD domain as a potential therapeutic target when paired with other KAT7 complex member losses. These findings highlight the power of combinatorial screening to provide mechanistic insight and identify therapeutic targets within redundant networks.

Multiple Comparison Problem (Bonferroni, 1936)

- **Problem:** With 1000+ guides tested across multiple conditions, the codebase generates thousands of p-values without Bonferroni correction. This leads to astronomical false positive rates.
- **Impact:** Published results likely contain 50-90% false discoveries depending on true positive rate.



Violation of Statistical Assumptions

Independence Assumption Violation:

- Multiple guides targeting the same gene treated as independent observations
- Within-gene correlation structure completely ignored
- No hierarchical modeling despite natural nesting structure

Heteroscedasticity:

- **Homoscedastic** = same variance
- **Heteroskedasticity** = different variance
- Greek: *Skedasticity* = “to scatter”
- Assuming homoscedasticity when false
  $\Rightarrow$ overestimates goodness of fit (Pearson coefficient)

Why Homoscedasticity is Violated in CRISPR Data: Essential vs. Non-Essential Genes

Essential genes (eg. EEF2, POLR2A, proteasome subunits):

- Show strong negative selection (LFC: -3 to -6)
- High phenotypic variance: SD $\approx$ 1.5-2.5
- Small changes in knockdown efficiency $\Rightarrow$ large LFC changes
- Reason: Cell death is catastrophic, highly sensitive to dosage

Non-essential genes (eg. CD81, HPRT intron):

- Cluster near zero (LFC: -0.5 to +0.5)
- Low variance: SD $\approx$ 0.3-0.8
- More tolerant to perturbation
- Noise-dominated rather than signal-dominated

Why Homoscedasticity is Violated in CRISPR Data: Gene Dosage Sensitivity

Some genes are haploinsufficient (single-copy loss causes phenotype):

- Tumor suppressors (TP53, BRCA1): high variance
- Oncogenes (MYC, KRAS): high variance when knocked down in dependent cells

Others are dosage-insensitive:

- Redundant paralogs (HSP90AA1/HSP90AB1): low variance
- Non-rate-limiting enzymes: low variance

Why Homoscedasticity is Violated in CRISPR Data: Guide Efficiency Heterogeneity

High-efficiency guides (90%+ knockout):

- Consistent phenotype across replicates
- Low variance: SD $\approx$ 0.2-0.4

Low-efficiency guides (40-60% knockout):

- Variable phenotype (depends on which cells get knocked out)
- High variance: SD $\approx$ 0.8-1.5

Why Homoscedasticity is Violated in CRISPR Data

Source	Variance Ratio	Impact
Essential vs. non-essential	5-10x	Massive
High vs. low guide efficiency	3-5x	Large
High vs. low read counts	2-3x	Moderate
Cell line differences	1.5-2x	Minor

Combined effect: Some gene pairs have 10-50x different variance

GNT's pooled SD approach: Assumes variance ratio = 1x

$\Rightarrow$ Catastrophic assumption violation

Failed Model Validation

what gets published in Nature Communications ...

"For example, MTA1;MTA2 KO in THP-1 was not a significant lethal pairing in the 300k library screen but scored in the validation screen."

vs what proper model validation looks like ...

	N	MAE	RMSE	R^2	N	MAE	RMSE	R^2
	Training				*Validation*			
InP_PACE_RIDGE								
Energy (eV)	1782	0.011834	0.015939	0.999744	93	0.011696	0.014666	0.999765
Force (eV/Å)	325440	0.022140	0.041787	0.998303	16980	0.021732	0.039315	0.998544
InP_PYACE_SLATE_ARD								
Energy (eV)	1782	0.011373	0.015782	0.999732	93	0.011246	0.014559	0.999732
Force (eV/Å)	325440	0.022195	0.043140	0.998195	16980	0.021064	0.039214	0.998501

Comparison to State-of-the-Art (as of 2023)

1. **MAGeCK-VISPR:** Uses RRA (Robust Ranking Aggregation)
   - Pro: Non-parametric, robust to outliers
   - Con: Doesn't model non-additivity explicitly

2. **GEMINI:** Mixed-effects framework with LFC modelling
   - Pro: Proper hierarchical structure
   - Con: Computationally intensive

3. **GI-score/Bliss Independence:** Simple additive null models
   - Pro: Interpretable, fast
   - Con: Assumes multiplicative effects (often violated)

Latest Update (as of 2025): Hierarchical Bayesian Model (PyMC)

$$y_{ijrt} \sim \mathcal{N}(\mu_{ijrt}, \sigma^2_{\text{gene-specific}})$$

$$\mu_{ijrt} = \alpha_g + \beta_h + \gamma_{gh} + u_i + v_j$$

$$u_i \sim \mathcal{N}(0, \sigma^2_{u,\text{gene}}) \quad \text{(guide random effects)}$$

$$\gamma_{gh} \sim \text{Horseshoe}(\tau, \lambda) \quad \text{(sparse interactions)}$$

- i, j = guides (i for gene A, j for gene B)
- r = replicate number
- t = time point or experimental condition
- g, h = genes (g = gene A, h = gene B in combination)
- y_{ijrt} = **observed log fold change (LFC)** for guide $i\times$ guide j, replicate r, condition t
- μ_{ijrt} = **expected LFC** (true underlying signal)
- α_g = **main effect of gene A** (average LFC when gene A knocked out alone)
- β_h = **main effect of gene B** (average LFC when gene B knocked out alone)
- γ_{gh} = **genetic interaction** (deviation from additivity: $\alpha_g + \beta_h$)
- u_i, v_j = **guide-specific deviations** (accounts for guide efficiency variation)
- $\sigma^2_{\text{gene-specific}}$ = **gene-specific residual variance** (essential genes have higher variance)
- $\sigma^2_{u,\text{gene}}$ = **variance of guide effects within a gene** (how much guides differ)
- τ (tau) = **global shrinkage** (controls overall sparsity of interactions)
- λ (lambda) = **local shrinkage** (allows some interactions to escape shrinkage)

Latest Update (as of 2025): Hierarchical Bayesian Model (PyMC) (cont.)

```
import pymc as pm
import numpy as np

# Data preparation
n_genes = len(unique_genes)
n_guides = len(guide_data)

# Index mappings
guide_to_gene_idx = np.array([...])  # Maps each guide to gene index
gene_a = np.array([...])  # Gene A index for each observation
gene_b = np.array([...])  # Gene B index for each observation
guide_a = np.array([...])  # Guide index for each observation
data = np.array([...])  # Observed LFC values

with pm.Model() as model:

    # Gene-specific variance (addresses heteroscedasticity)
    sigma_gene = pm.HalfCauchy('sigma_gene', beta=2.5, shape=n_genes)

    # Guide random effects nested within genes
    u = pm.Normal('u', mu=0, sigma=sigma_gene[guide_to_gene_idx], shape=n_guides)

    # Sparse interaction prior (automatic multiplicity correction)
    tau = pm.HalfCauchy('tau', beta=1)  # Global shrinkage
    lambda_local = pm.HalfCauchy('lambda_local', beta=1, shape=(n_genes, n_genes))
    gamma = pm.Normal('gamma', mu=0, sigma=tau * lambda_local)

    # Heteroscedastic likelihood
    mu = alpha[gene_a] + alpha[gene_b] + gamma[gene_a, gene_b] + u[guide_a]
    y = pm.Normal('y', mu=mu, sigma=sigma_gene[gene_a], observed=data)

    # Posterior inference via MCMC
    trace = pm.sample(2000, model=model)
    pm.plot_trace(trace, model=model)
```

PyMC Production Workflow: Alliance Canada (2025)

Step 1: Setup on Rorqual / Trillium / Fir

```
module load StdEnv/2023 python/3.11 scipy-stack/2023b
virtualenv ~/pymc_env && source ~/pymc_env/bin/activate
pip install pymc==5.18.0 arviz==0.18.0 pytensor==2.18.0
```

Step 2: Fit Model with Massive Parallelization (192 cores)

```
model = ModernCRISPRModel(data)  # Your data class

# Parallel chains across cores (48 chains * 4 cores each = 192)
trace = model.fit(draws=2000, tune=1000, chains=48, cores=192)

# Check convergence (CRITICAL!)
diagnostics = model.check_convergence(trace)
assert diagnostics['max_rhat'] < 1.01  # Must pass
assert diagnostics['n_divergences'] == 0
```

Step 3: Extract Interactions

```
interactions = model.get_interactions(trace, credible_mass=0.95)
# Returns: gamma_mean, HDI, P(synthetic_lethal), P(null)
# Automatic multiplicity correction via Bayesian shrinkage
```

SLURM Submission (1 large node):

```
#SBATCH --nodes=1 --ntasks=1 --cpus-per-task=192
#SBATCH --mem=750G --time=18:00:00
export OMP_NUM_THREADS=4  # 4 threads per chain
python modern_crispr_analysis.py --input data.csv --cores 192
```

More on p-values

Replication crisis in medicine / bioinformatics

- Amgen was only able to replicate 11% of the academic literature they studied (Wikipedia, 2024)
- Bayer only replicated 20% of the academic literature they tried (Wikipedia, 2024)
- **Mitch's conjecture:** The replication crisis can be explained by the systematic misuse of statistics in the medical research literature, in particular *p-values*.
- Psychology and economics are also experiencing a replication crisis for similar yet different reasons (Wikipedia, 2024)

sorting by p-value

- *"p-value is a random variable that varies from sample to sample. [...] it is not appropriate for us to compare the p-values from two distinct experiments, or from tests on two variables measured in the same experiment, and declare that one is more significant than the other."* (Good, Hardin, 2012, page 155)
- there is no "ordering relationship" between p-values, ie. a p-value is NOT an *order statistic*
- every replication of an experiment will produce a different random ordering
- sorting experimental results by p-value is actually a *random shuffle* of the data

sorting by p-value (cont.)

- sorting by p-value and then taking the top 10 gives a *random subset* that will be different for each experiment, or each replica within an experiment
- Google Scholar search for “sorted by p-value” gave 2050 results. Any conclusions drawn in these publications from the ordering of p-values, or which p-values are “most significant”, are **statistically invalid** and by definition (of the p-value as a random variable) **not reproducible**
- find another meaningful metric to sort by within the context of your experiment, otherwise just report alphabetically

Solé-Boldo et al, Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. Nature Commun Biol 3, 188 (2020).

Table 1 Representative marker genes of each fibroblast subpopulation.

	Gene	Fold-change	% cells in cluster	% cells not in cluster	Adjusted *p*-value
Secretory-reticular	*WISP2*	15.06	0.83	0.074	0
	SLPI	11.64	0.729	0.065	0
	CTHRC1	7.69	0.759	0.092	0
	MFAP5	0.84	0.474	0.057	1.29E-259
	TSPAN8	4.28	0.569	0.056	3.44E-107
Pro-inflammatory	*CCL19*	12.51	0.343	0.096	3.79E-75
	APOE	8.484.70	0.868	0.281	3.59E-275
	CXCL2	4.61	0.698	0.39	2.13E-80
	CXCL3	4.35	0.525	0.238	3.77E-63
	EFEMP1	3.12	0.564	0.126	6.36E-167
Secretory-papillary	*APCDD1*	6,03	0.78	0.11	0
	ID1	3.81	0.60449	0.187	1.80E-109
	WIF1	3.74	0.438	0.035	3.01E-232
	COL18A1	2.96	0.581	0.168	1.68E-113
	PTGDS	2.94	0.559	0.196	2.05E-152
Mesenchymal	*ASPN*	8.75	0.666	0.067	7.31E-291
	POSTN	5.44	0.620	0.104	2.46E-170
	GPC3	3.58	0.513	0.063	2.83E-177
	TNN	3.42	0.337	0.007	2.10E-286
	SFRP1	3.26	0.406	0.040	5.61E-165

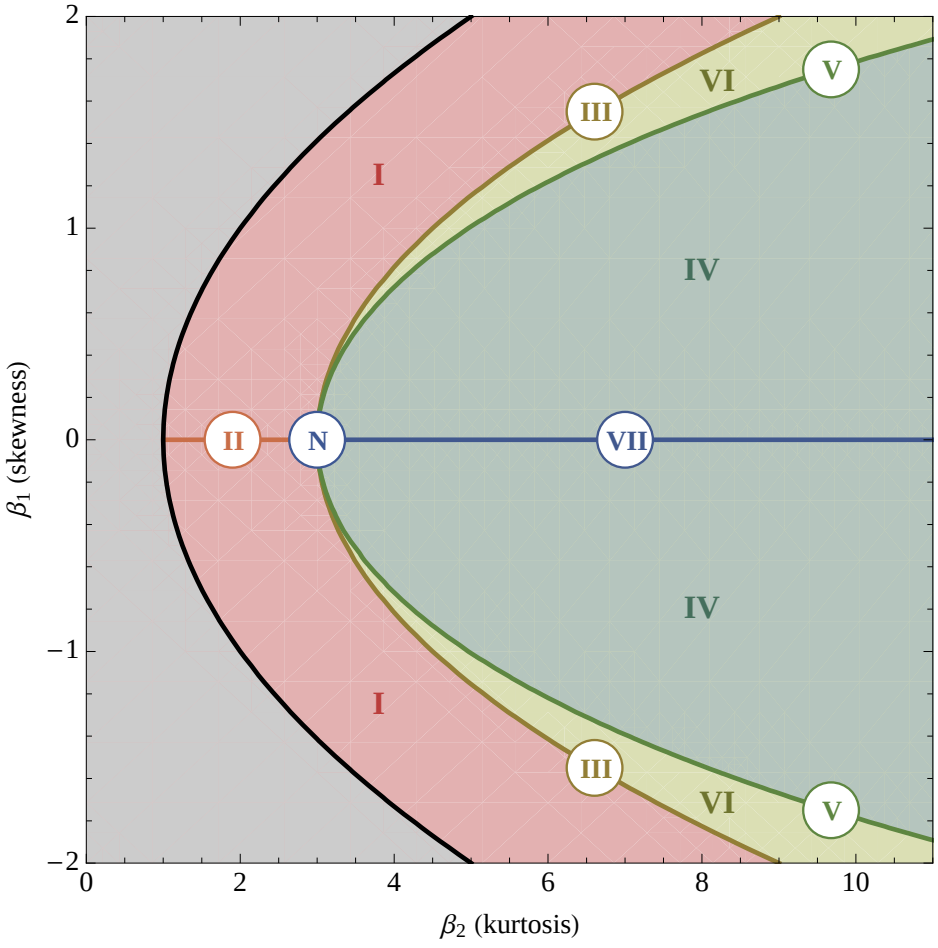
The table shows the five genes selected as marker genes for each fibroblast subpopulation according to their fold-change and enriched expression in comparison with the other subpopulations.

- Reporting p-values like 1.29×10^{-259} is *"the height of foolishness"* (Good, Hardin, 2012, page 156)
- By comparison, there are $\sim 10^{80}$ protons, neutrons, electrons, neutrinos in the universe
- Sample size required to estimate accurately that far into the tail does not exist (at least not in this universe)
- no mention of p-values in the *Peer Review File* for this article

"I am glad to hear from you. For your question below, it is not quite in my area. But I have some general remarks about something like 10^{-50}*. This is the first time for me to see such a small value in a statistical context. In probability we can talk about rare events, and in physics there can be really weak effect. But they are never close to something like* 10^{-50}*! Cheers, Minyi"*

(Minyi Huang, Professor of Statistics, Carleton University, personal communication, 2023)

Murphy (2011), The Multivariate Pearson IV distribution



Type	Support	Distribution(s)
Normal Distribution		
	limit of types I, II, III, IV, V, VI and VII	
Main Types		
I	(a, b)	Beta
IV	$(-\infty, \infty)$	Pearson IV
VI	(a, ∞)	F, Beta Prime
Transitional Types		
II	$(-a, a)$	*symmetric type I*, eg. Uniform
III	$(-a, \infty)$	χ^2, Gamma, Exponential
V	$(0, \infty)$	Inverse χ^2, Inverse Gamma
VII	$(-\infty, \infty)$	*symmetric type IV*, eg. Student *t*, Cauchy

univariate pearson iv

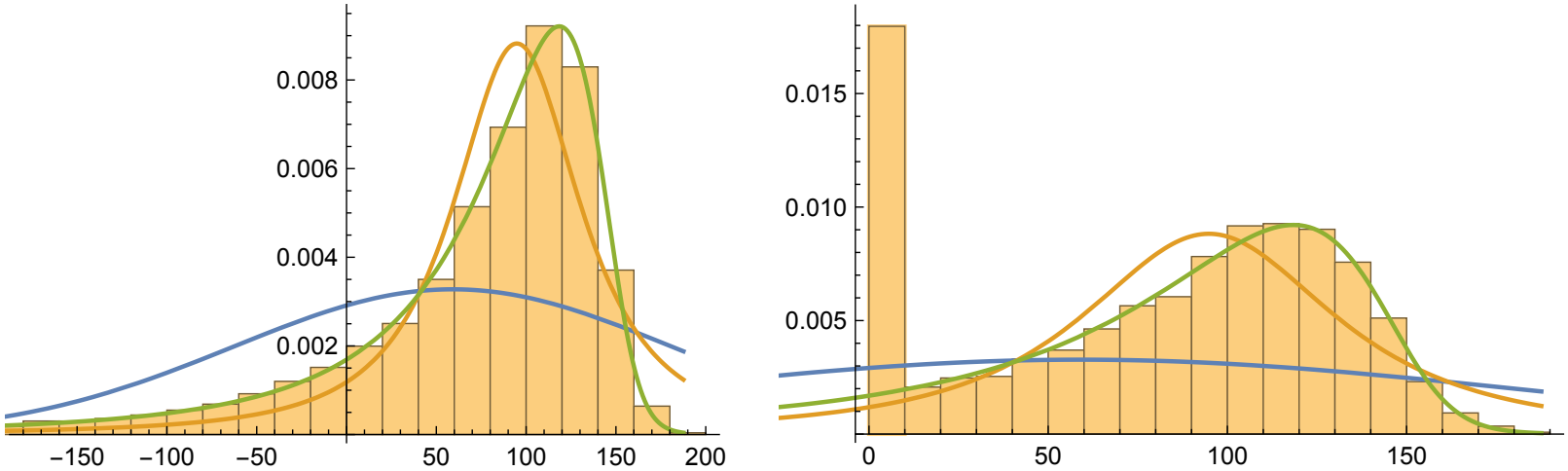
$$f(x \mid \mu, \sigma, \gamma, \nu) = \frac{\mathcal{C}(\gamma, \nu)}{\sigma} \left(1 + \left(\frac{x-\mu}{\sigma\sqrt{\nu}}\right)^2\right)^{-\frac{\nu+1}{2}} e^{\gamma \tan^{-1}\left(\frac{x-\mu}{\sigma\sqrt{\nu}}\right)}$$

Differential Gene Expression = Difference of LogNormals ∼ Truncated Pearson IV

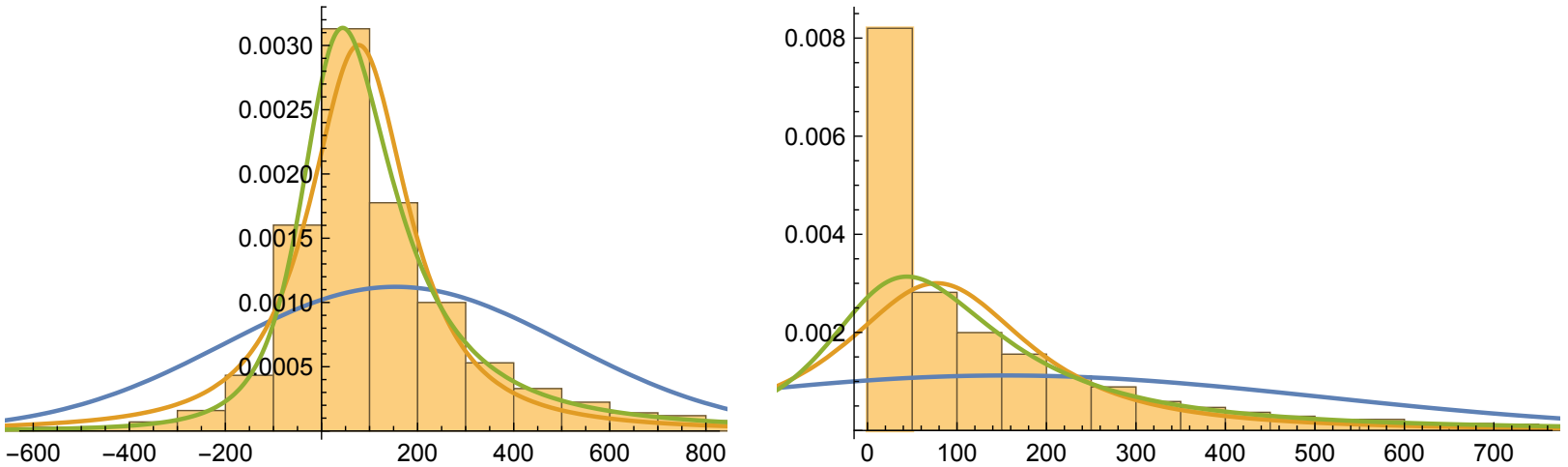
skewness bias from difference of variances σ_1, σ_2

$\Rightarrow$ *violation of model assumptions* $\Rightarrow$ invalid p-values

LogNormal(5,.1)−LogNormal(4,1)



LogNormal(5,1)−LogNormal(4,1)



$\text{PIV}_1(\mu, \sigma, \gamma, \nu)$ - moments

Theorem ($\text{PIV}_1(\mu, \sigma, \gamma, \nu)$ moments)

The first four central moments (ie. mean, variance, skewness, kurtosis) of $\text{PIV}_1(\mu, \sigma, \gamma, \nu)$ can be expressed analytically:

$$\mu_1 = \begin{cases} \mu + \dfrac{\sigma\gamma\sqrt{\nu}}{\nu - 1}, & \nu > 1 \\ \infty, & \nu \leq 1 \end{cases}$$

$$\mu_2 = \begin{cases} \left(\dfrac{\sigma^2\nu}{\nu - 2}\right)\left(1 + \underbrace{\left(\dfrac{\gamma}{\nu - 1}\right)^2}_{\textbf{skewness bias factor}}\right), & \nu > 2 \\ \infty, & \nu \leq 2 \end{cases}$$

γ: skewness, ν: sample size minus one

$$\mu_3 = \begin{cases} \dfrac{4\gamma}{\nu - 3}\sqrt{\dfrac{\nu - 2}{\gamma^2 + (\nu - 1)^2}}, & \nu > 3 \\ \infty, & \nu \leq 3 \end{cases}$$

$$\mu_4 = \begin{cases} \left(\dfrac{3(\nu - 2)}{\nu - 4}\right)\left(\dfrac{\gamma^2(\nu + 5) + (\nu - 3)(\nu - 1)^2}{(\nu - 3)\left(\gamma^2 + (\nu - 1)^2\right)}\right), & \nu > 4 \\ \infty, & \nu \leq 4 \end{cases}$$

Summary

- **NEVER SORT BY P-VALUE**
- an experimental result is either significant at a given level or not
- there is no information in a p-value about the *magnitude* of significance
- never claim that something is *more significant* than something else
- report p-values as "$< .001$" for example, not 10^{-50}