Interpretation of *P*-values: Challenges for the Replication and Comparison of Statistical Results

Ilana Belitskaya-Lévy, Ph.D.
Cooperative Studies Program Coordinating Center
VA Palo Alto Health Care System

VA Statisticians' Association

October 29, 2013

Acknowledgments

- Joint work with:
  - Laura Lazzeroni (Stanford University) and
  - Ying Lu (VA Palo Alto and Stanford University)

- Funding:
  - VA Palo Alto Cooperative Studies Program-Wide DNA Bank (CSP#478)
  - NIH/NIMH R01 MH086135 and the Consortium on the Genetics of Schizophrenia (COGS)

Outline

- Genome-wide association studies (GWAS)
  - Genome-wide significance
  - Failure to replicate, winner's curse
  - Follow-up ("replication") studies
- *P*-values
  - Sampling variability of a *p*-value
  - How likely is replication?
  - How confident are we in the relative meaning of two different *p*-values*?*
- Implications and conclusions

Poll Question #1

- What is your primary role in VA?
  - biostatistician or statistical programmer
  - student, trainee, or fellow
  - clinician
  - researcher
  - manager or policy-maker
  - other

Genome-wide significance

A single modern genomics study may include a million or more statistical tests.

To reduce the number of false positive results, very strict significance levels are used.

Thus, significance levels on the order of 10^{-8} have replaced the traditional 0.05 value.

This talk examines the unintended and unappreciated consequences of this change.

Concerns regarding GWAS studies

1. Failure to replicate initial findings

2. Even when replicated, strength of genetic association vary from study to study

Why?

Genetic heterogeneity?

Winner's curse?

Publication bias?

Genomic “Replication” Studies

- Genetic studies often include an initial large study followed by a more targeted follow-up (“replication”) study within the same publication.
- Why? Initial results may be disappointing or ambiguous given stringent genome-wide testing criteria.
- Additionally, many journals (e.g. NEJM) require replication BEFORE they will publish a genetic association due to skepticism regarding genetic associations.

How Are Genetic Variants Selected for Replication?

- Custom-designed genotyping chips (smaller & cheaper than the original) have a pre-designated amount of "real estate."
- Selection criteria include:
  - Small *p*-values (< 0.001 or <0.0001)
  - Large estimated effect sizes (odds ratios)
  - Previous reports of association
  - Biologically relevant pathways
  - Plausible function of the variant

  - Or a combination of the above

Designing Follow-up or Replication Studies

- Investigators often use an *ad hoc* approach to design follow-up studies.

  Ignoring important factors such as the winner's curse.

- In this talk, we provide explicit numerical results that show that even extremely small *p*-values provide surprisingly little information as to the likelihood of subsequent replication or the relative importance of underlying association.

Questions

1. What is the sampling variability of a *p*-value?

2. How likely is an independent study to replicate a given *p*-value?

3. How confident can we be that a more significant p-value reflects a larger underlying effect than a less significant *p*-value?

The Key Assumption

We are using a test statistic, Z, with an asymptotic Normal distribution. In the usual notation,

$$Z = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \sim N(0,1)$$

Thus, results apply in most common statistical settings.

Genomics terminology is used for convenience:

GWAS = genome-wide association studies

SNP = single nucleotide polymorphisms

Question 1:
What is the sampling variability of a *p*-value?

P-value is an informative measure of strength of statistical results

P-values are also highly variable random statistics of the data, with sampling distributions of their own.

Question 1:
What is the sampling variability of a *p*-value?

- 1,000 Simulated studies of the same SNP

- For each simulation study:

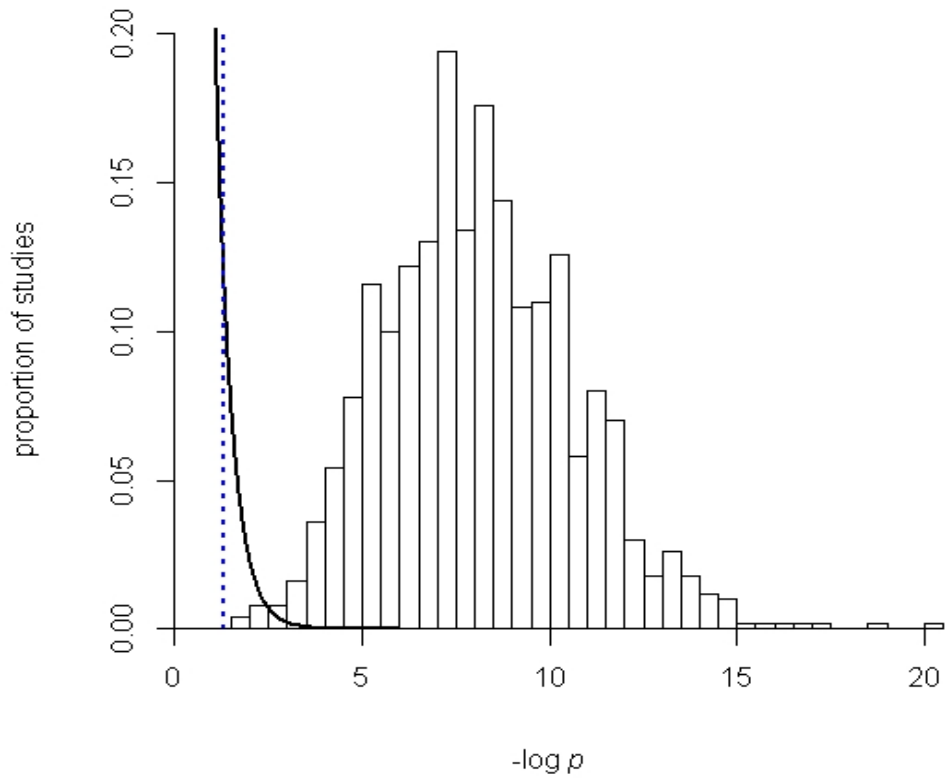
2000	Sample size
1.74	Odds ratio
30%	Minor allele frequency
10%	Major homozygote disease risk

- -log *p* (base 10) from a logistic regression analysis

1000 simulated studies of same SNP

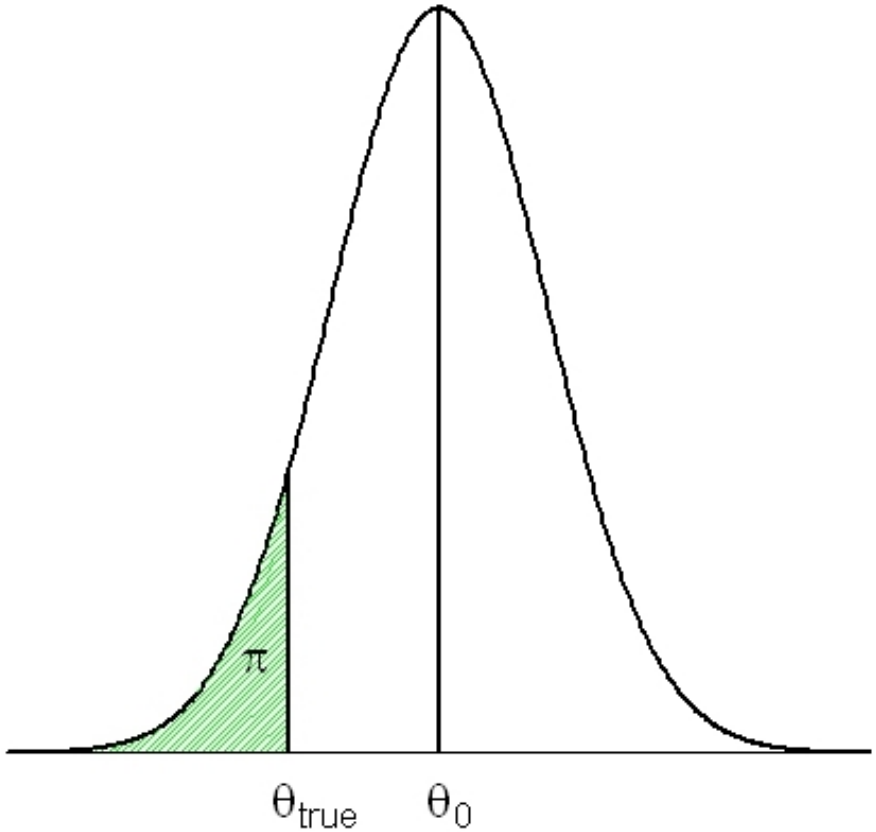
Median *p*: 10^{-8}
Min. *p*: 5.6 10^{-21}
Max. *p*: 0.015

In real study, we see one value of *p*, but do not know distribution or median relative to *p*



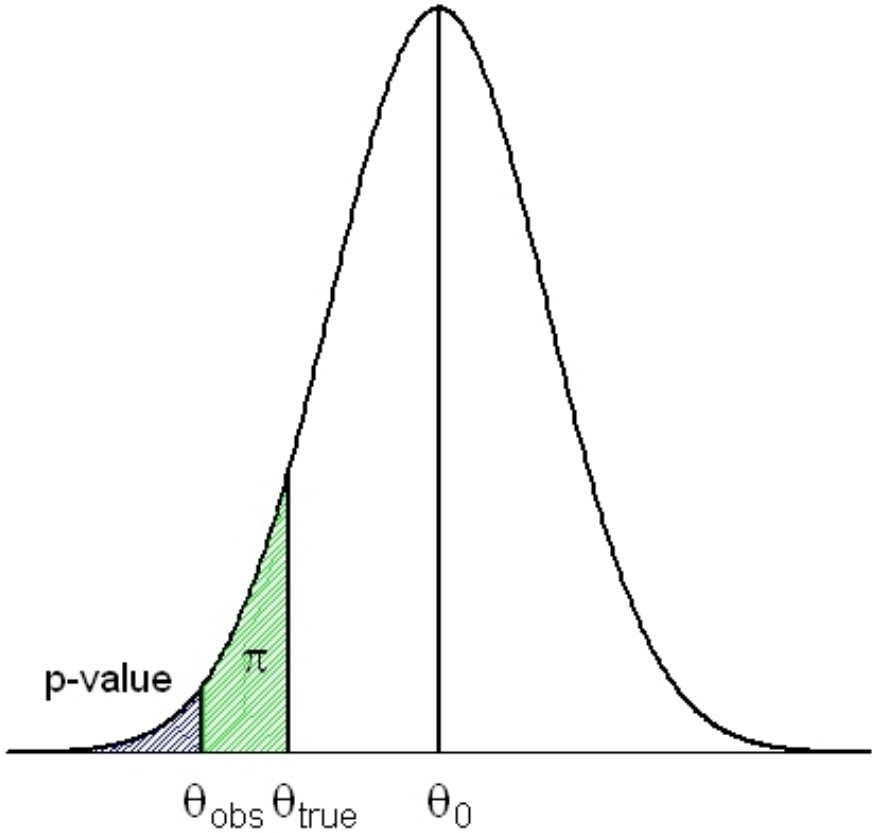
Vertical line at *p* = 0.05
Solid curve = null exponential density

Does *p*-value estimate anything?



Yes. We define π or the π-value to be the value of *p* when the estimate (i.e. test statistic) exactly equals true, unknown θ.

Does *p*-value estimate anything?



Yes. We define π or the π-value to be the value of *p* when the estimate (i.e. test statistic) exactly equals true, unknown θ.

Confidence interval for π

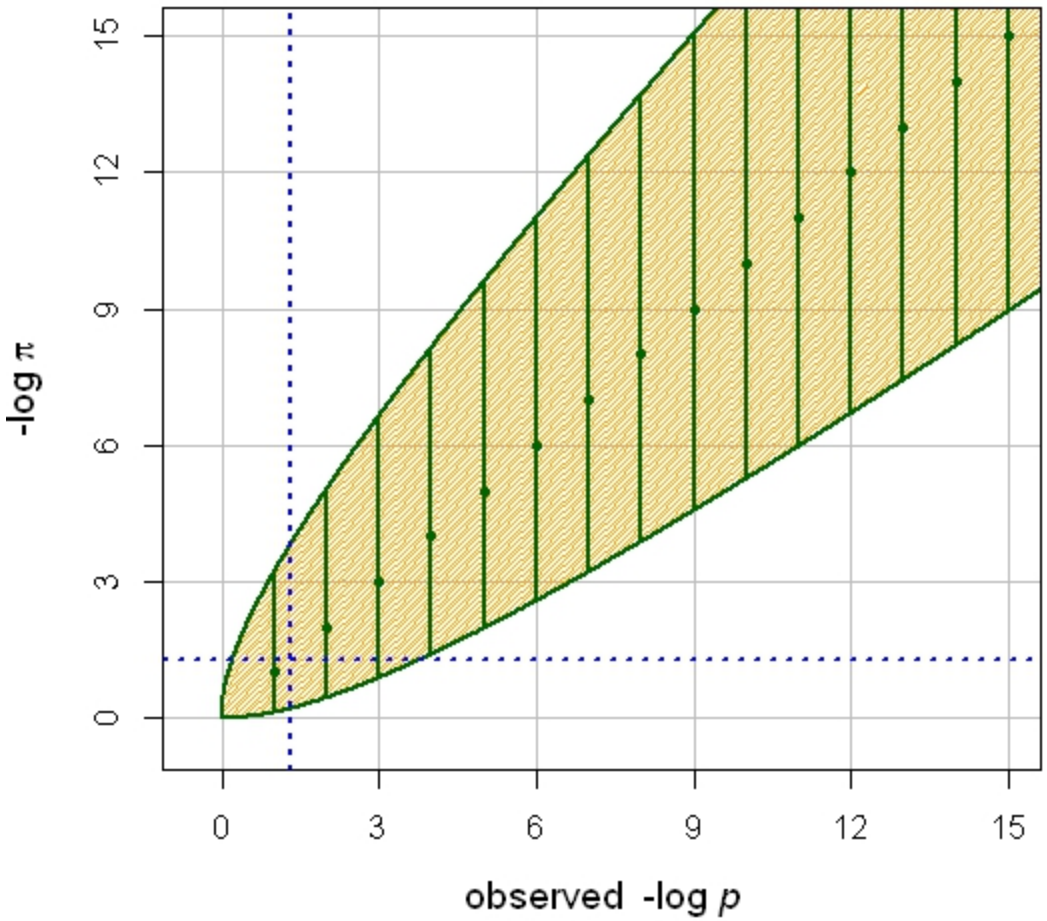
- Observe p_1 in the initial study
- Calculate $Z_1 = Z(p_1)$
- A 95% confidence interval for $Z(\pi)$ is:

$$Z(\pi) \in Z(p) \pm 1.96$$

- The 95% confidence interval for π is:

$$\pi \in [\, p(Z(p) + 1.96),\ p(Z(p) - 1.96)]$$

Confidence Intervals for π



95% confidence intervals apply to p-values in any large-sample context

X-axis: -log p_1

Y-axis: -log π

Confidence intervals extend vertically from top to bottom sides of the green funnel.

Question 2:
How likely is replication?

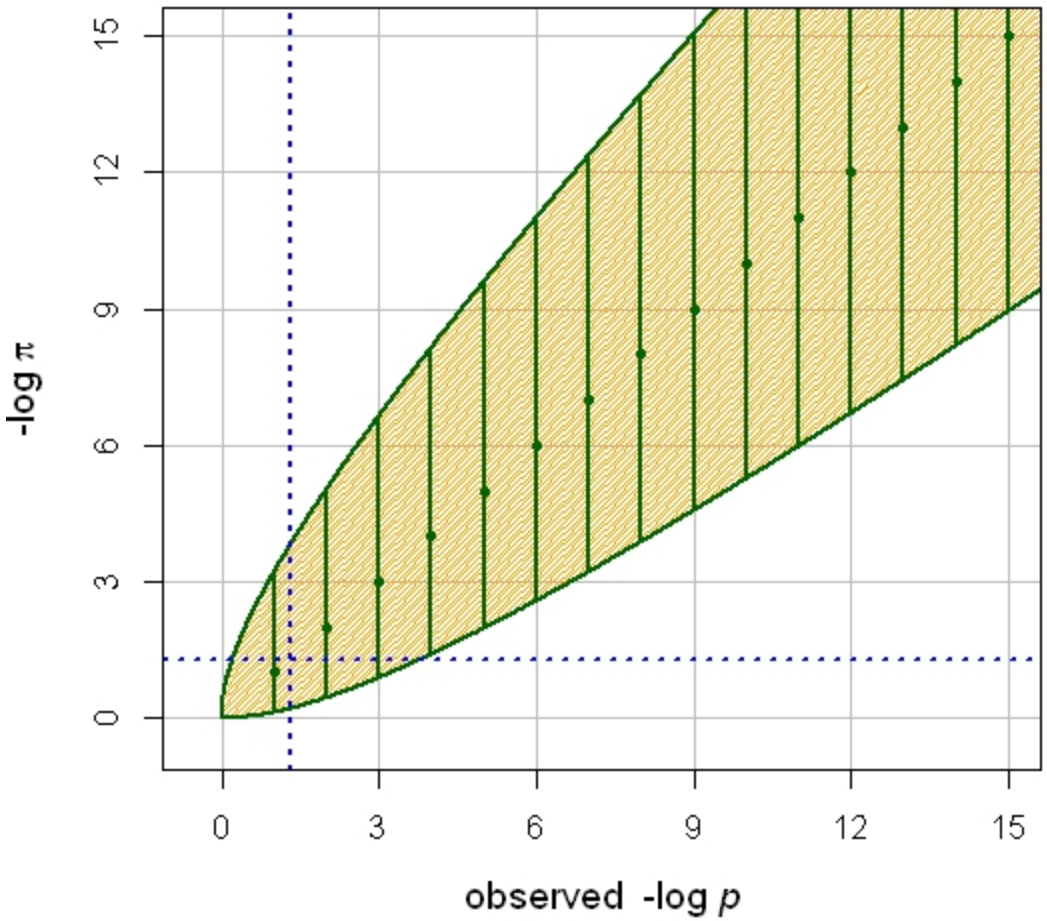
We construct prediction intervals for p_2 from a replication study given p_1 in an initial study.

For example, let n_1, n_2 be the original & replication sample sizes. A 95% prediction interval for Z_2 is

$$\sqrt{\frac{n_2}{n_1}}Z(p_1) \pm 1.96\sqrt{1+\frac{n_2}{n_1}}$$

This can be converted to a prediction interval for p_2.

Results for Equal Study Sizes



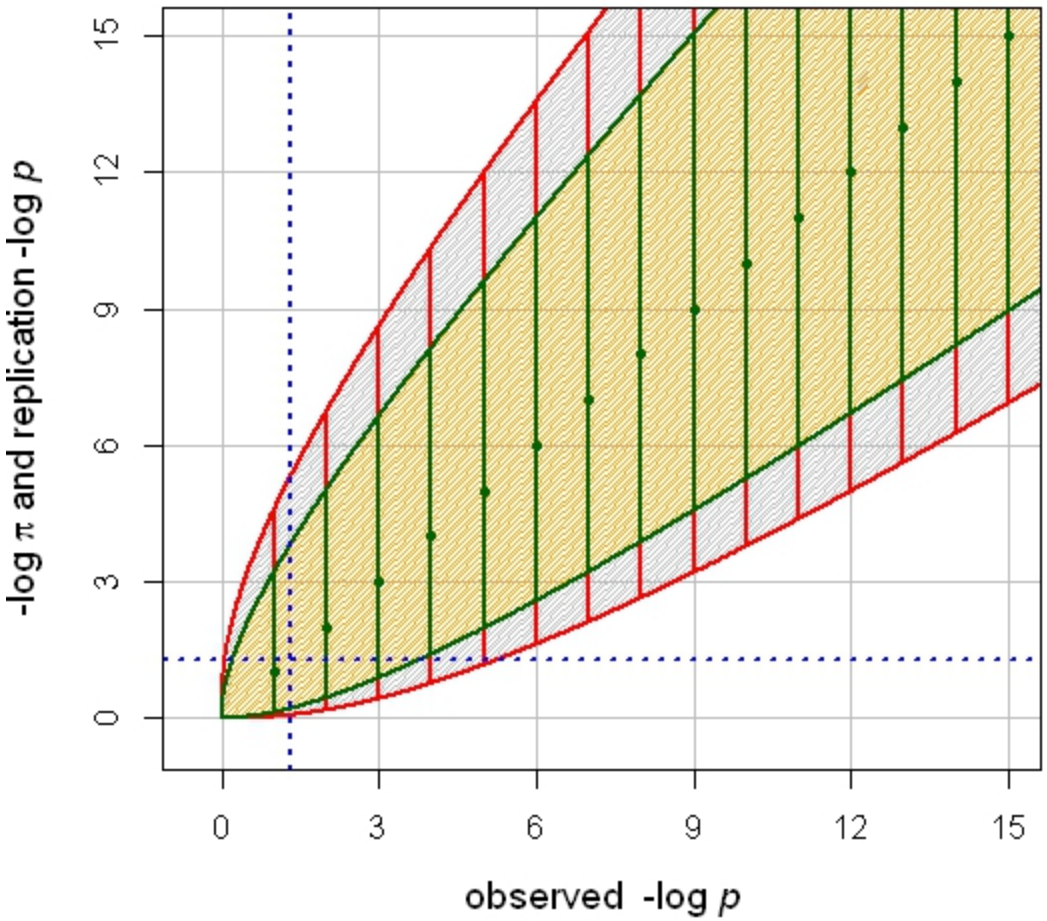
95% confidence intervals apply to p-values in any large-sample context

X-axis: -log p_1

Y-axis: -log π

Confidence intervals extend vertically from top to bottom sides of the green funnel.

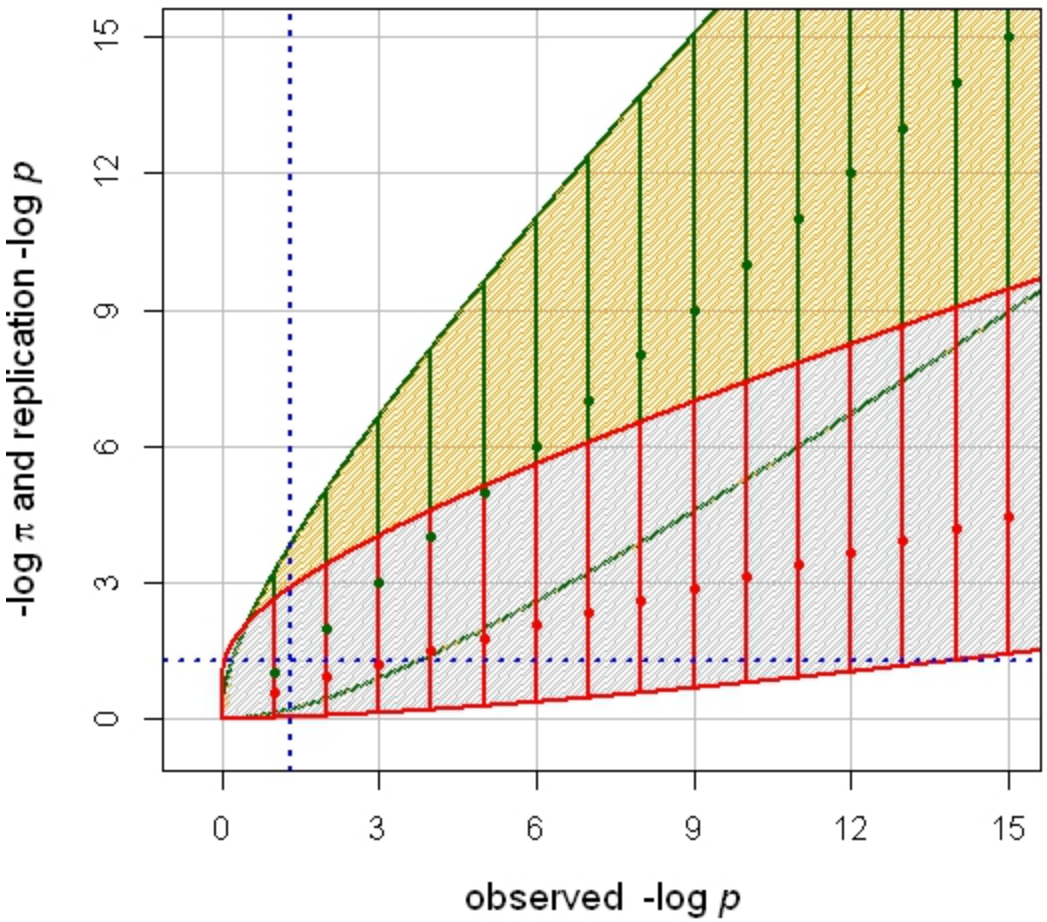
Results for Equal Study Sizes



95% prediction intervals extend vertically from top to bottom of outer red funnel.

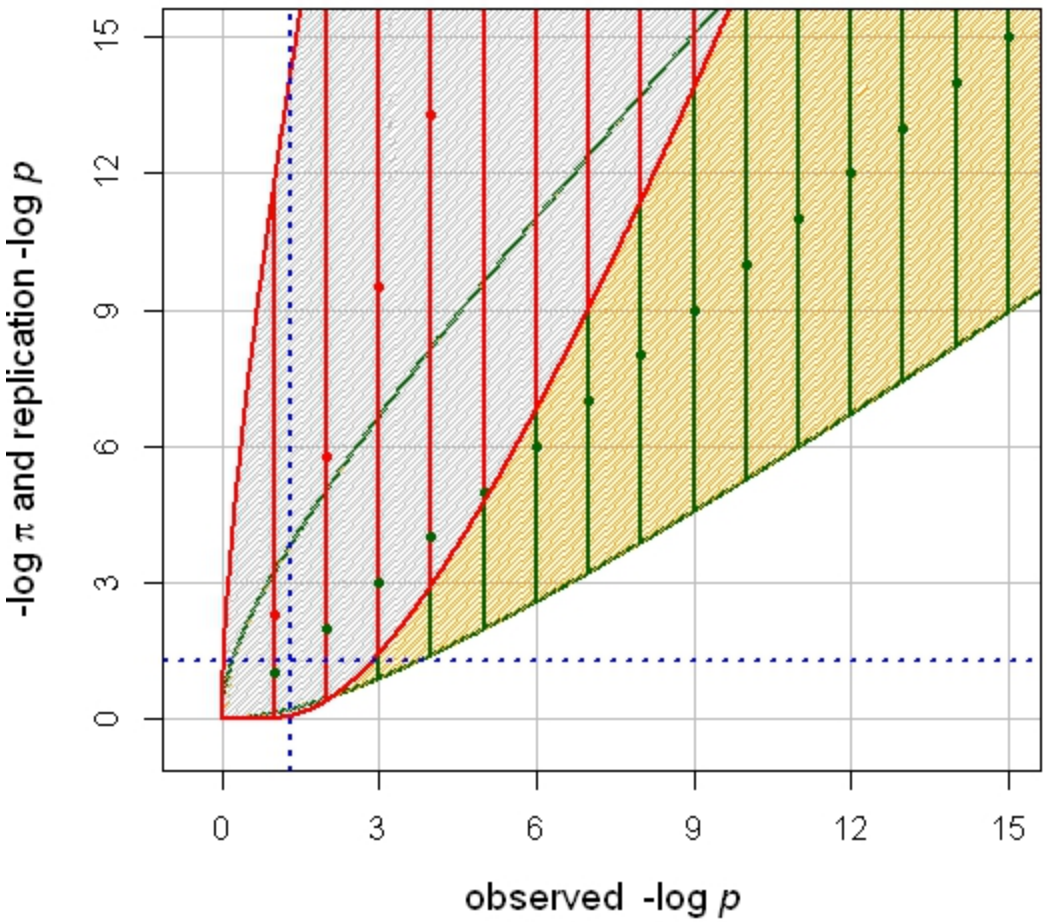
Y-axis: -log p_2 predicted for replication.

Unequal study sizes shift π & the predicted *p*



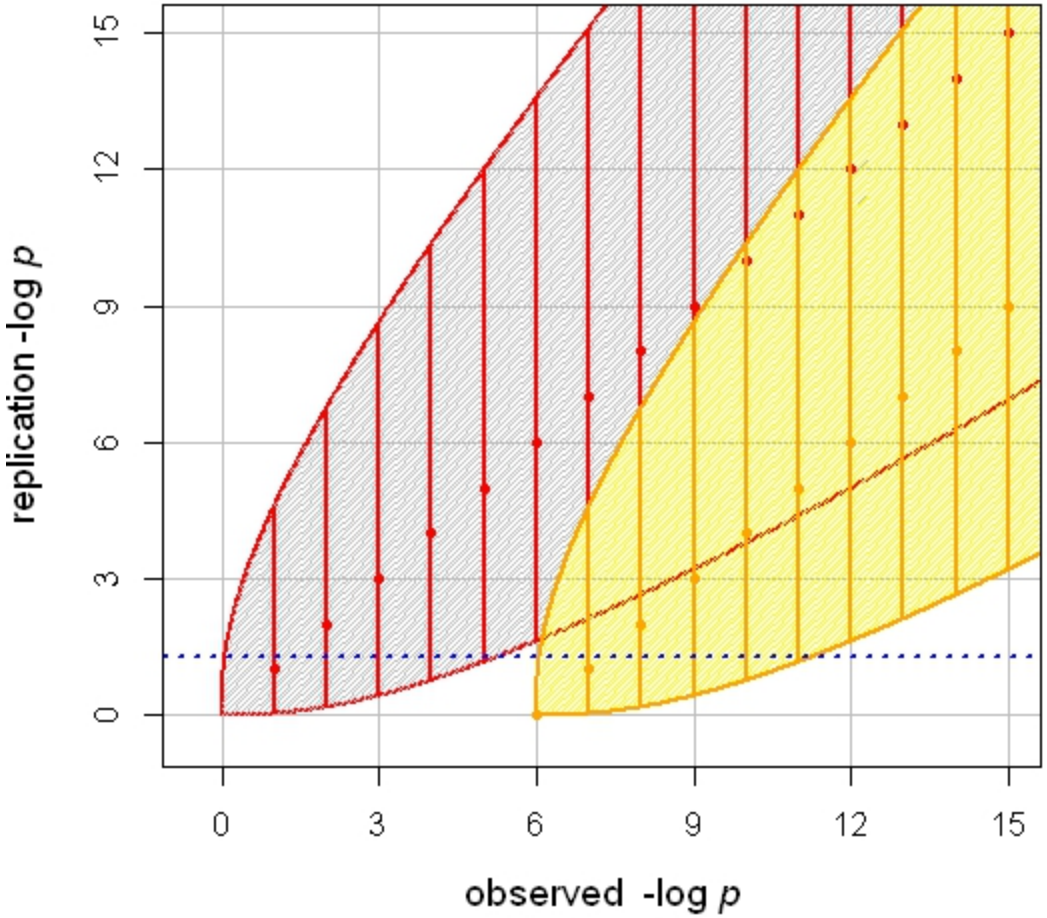
Here, the replication study is 1/4 the size of the original study.

Unequal study sizes shift π & the predicted *p*



Here, the replication study is 4 times the size of the original study.

Effect of accounting for multiple comparisons.



1M tests in original study

1 test in replication study

Bonferroni correction for 1M tests in original study.

No correction in replication study.

Meta-analysis

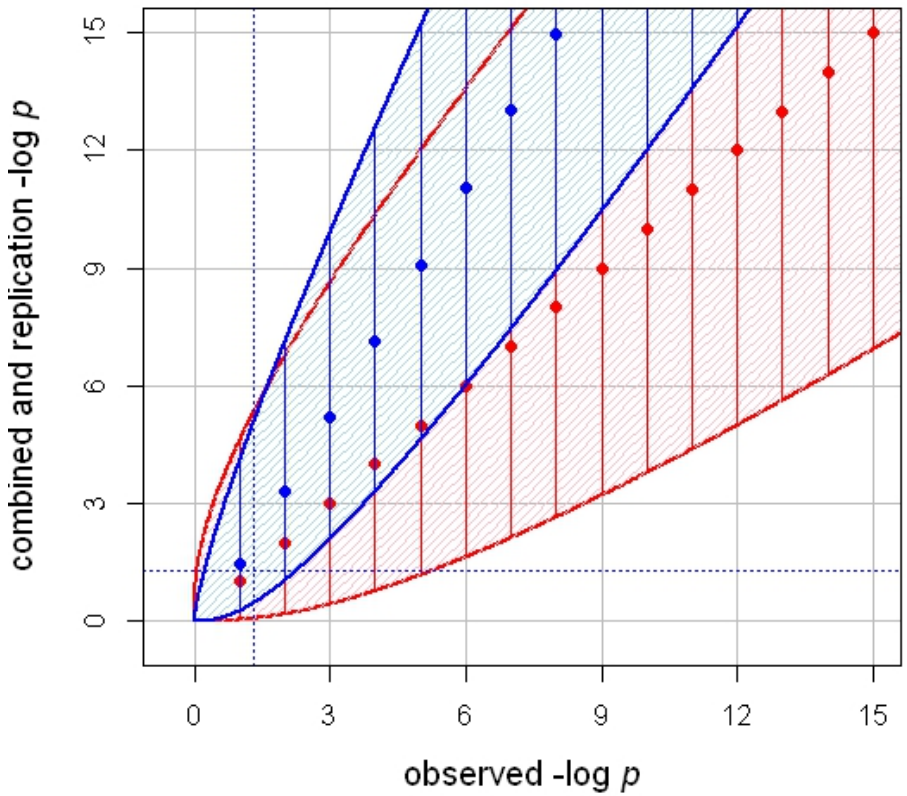
- Suppose a sample of size n_1 has been analyzed.
- Question: replicate in a sample of size n_2 or augment the original sample and use a combined analysis?

$$Z_{combined} = \frac{\sqrt{n_1} Z_1 + \sqrt{n_1} Z_2}{\sqrt{n_1 + n_2}}$$

- 95% prediction interval for combined *Z*:

$$Z(p_1)\sqrt{1+\frac{n_2}{n_1}} \pm 1.96\sqrt{\frac{n_2}{n_1}}$$

Replication vs Combined Analysis



Equal study sample sizes

<u>Blue:</u> prediction interval for $Z_{combined}$

<u>Red:</u> prediction interval for replication p-value

Question 3: How confident are we in the relative meaning of two different *p*-values*?*

Suppose we observe *p*-values for two different SNPs.

We ask "Is the effect of the more significant SNP really bigger than the effect of the less significant SNP?"

For example, which of the two SNPs to select for a follow-up study?

SNP comparison

SNP 1: $H_0 : \theta_1 = 0$ $H_a : \theta_1 > 0$

SNP2: $H_0 : \theta_2 = 0$ $H_a : \theta_2 > 0$

Observe *p*-values p_1 and p_2 such that $p_1 < p_2$

How confident are we that the estimated order is the "correct order"?

We want to construct a test of $\theta_1 = \theta_2$

What to compare?

$$H_0: \theta_1 = \theta_2$$

E.g.

- the log odds ratios
- the π-values, π_1 and π_2

Test Statistic

$$T_1 = \frac{Z(p_1) - r_1 Z(p_2)}{\sqrt{1 + r_1^2}}$$

$$r_1 = \frac{\sigma_2 / \sqrt{n_2}}{\sigma_1 / \sqrt{n_1}}$$

T_i is a N(0,1) test statistic of H_0: $\theta_1 = \theta_2$

If T is positive => stronger effect of SNP1
If T is negative => stronger effect of SNP2

Evidence Ratio

Let q_1 be p-value for a one-sided test with H_a: $\theta_2 > \theta_1$
Let q_2 be p-value for a one-sided test with H_a: $\theta_1 > \theta_2$

Evidence Ratio

q_1/q_2 in favor of SNP1 compared to SNP2

Evidence ratio of 95:5 means that we can conclude $\theta_1 > \theta_2$ in a one-sided , 5% significance test of H_0: $\theta_1 = \theta_2$.

Evidence ratio in favor of θ_1 vs θ_2

More significant	Log difference between p_1 and p_2						
-log p_1	0	1	2	3	4	5	6
10	50:50	60:40	70:30	79:21	87:13	93:7	97:3
9	50:50	61:39	71:29	81:19	89:11	95:5	98:2
8	50:50	61:39	73:27	83:17	91:9	96:4	99:1
7	50:50	62:38	75:25	85:15	93:7	98:2	100:0
6	50:50	64:36	77:23	88:12	96:4	99:1	
5	50:50	65:35	80:20	91:9	98:2		
4	50:50	67:33	84:16	96:4			
3	50:50	71:29	90:10				

r = 1

Without accounting for multiple comparisons

Implications and conclusions

- *P*-values vary greatly across studies, even when underlying effects, population and study design are identical.

- Studies can have good power to reject the null hypothesis of no association, while providing little information with respect to the reproducibility or relative strength of the true association.

- The more significant the *p*-value, the greater the variability.

Implications and conclusions

- Replication *p*-values can differ from an initially significant GWAS finding, especially after multiple testing correction.

- *P*-values provide little resolution to distinguish the true, relative importance of different SNPs.

Our results

- Support proposals of others to combine multiple lines of biological evidence in deciding which results to investigate further, rather than depending on association *p*-values alone.

- Can be used to design successful replication studies.
  - The prediction intervals depend on the relative size of the replication and initial study.
  - The prediction intervals require no assumptions about the unknown effect sizes or the initial sample size.

- Apply in most statistical settings, with or without multiple testing issues.

We gratefully acknowledge support from:

The VA Palo Alto Cooperative Studies Program DNA Bank

NIH/NIMH R01 MH086135 and the
Consortium for the Genetics of Schizophrenia (COGS)

Additional slides

P-value variability

In a real study, we see only one *p*-value and don't know where the rest of the distribution lies relative to it.

p (unlogged) is most variable under the null hypothesis.

log *p* is more variable under the alternative hypothesis.

Using log *p* instead of *p* affects our intuitive interpretation.

Confidence interval for π

- Consider a one-sided lower-tail hypothesis test for the population parameter θ:

$$H_0 : \theta \geq \theta_0$$
$$H_a : \theta < \theta_0$$

- Assume without loss of generality that $\theta_0 = 0$
- Given *p*-value, we can construct *Z* = *Z*(*p*)
- Given *Z*, we can find *p* = *p*(*Z*)
- *Z*(*p*) & *p*(*Z*) are functions mapping one to the other
  - *P(Z)* is the cumulative distribution function of the standard Normal distribution and *Z(p)* is its inverse.

Prediction interval for replication *p*-value

- Z_i is $N(\sqrt{n_i}\theta/\sigma,1)$
- Consider $R = Z_2 - \sqrt{n_2/n_1}Z_1$ which is Normal with expected value 0 and variance $1+n_2/n_1$
- Confidence interval for R is

$$\pm Z(1-\beta/2)\sqrt{1+n_2/n_1}$$

- Confidence interval for Z_2 is

$$\sqrt{\frac{n_2}{n_1}}Z(p_1) \pm 1.96\sqrt{1+\frac{n_2}{n_1}}$$

Coverage rate of the prediction interval

For each of the 1,000 *p*-values in Figure 1, we computed a separate 95% prediction interval without a multiple testing adjustment and calculated how many of the other 999 *p*-values it covered. The estimated coverage rate for our prediction interval procedure was 94.9%, consistent with theoretical results. The percentage of replication *p*-values covered varied across the distribution of initial *p*-values, ranging from 17.3% for $p = 0.015$ to 99.6% for $p = 10^{-8}$. It should be noted that these conditional coverage percentages would differ if the unknown effect size or sample size differed, even if the same initial *p*-value was observed.

Two-sided *p*-values

If the *p*-values are two-sided, dividing the observed value of p_1 yields the more significant of the two corresponding one-sided *p*-values.

The prediction interval for a one-sided *p*-value is then calculated and converted back to two-sided *p*-values using the equation

$$p_{two\text{-}sided} = 2 \min(p_{one\text{-}sided}, 1\text{-}p_{one\text{-}sided})$$

What to compare?

1. We can compare $E_1 = \theta_1$ to $E_2 = \theta_2$

   E.g. to compare the log odds ratios (must know relative ratio of standard errors, $\sigma_1 / \sqrt{n_1}$ and $\sigma_2 / \sqrt{n_2}$)

2. Or we can compare $E_1 = \theta_1/\sigma_1$ to $E_2 = \theta_2/\sigma_2$

   E.g. to compare R^2s from a linear regression model (need only n_1/n_2)

3. Or we can compare the π-values, π_1 and π_2

Test statistics

1. To compare θ_1 to θ_2, the test statistic is

$$T_1 = \frac{Z(p_1) - r_1 Z(p_2)}{\sqrt{1 + r_1^2}}$$

$$r_1 = \frac{\sigma_2 / \sqrt{n_2}}{\sigma_1 / \sqrt{n_1}}$$

2. To compare θ_1/σ_1 to θ_2/σ_2, the test statistic is

$$T_2 = \frac{Z(p_1) - r_2 Z(p_2)}{\sqrt{1 + r_2^2}}$$

$$r_2 = \sqrt{\frac{n_1}{n_2}}$$

3. To compare π_1 and π_2, the test statistic is

$$T_3 = \frac{Z(p_1) - Z(p_2)}{\sqrt{2}}$$

$$r_3 = 1$$

Evidence Ratio

- T_i is a N(0,1) test statistic of H_0: $E_1=E_2$, where

  $E_1= \theta_1$ or θ_1/σ_1 or $1-\pi_1$

  $E_2= \theta_2$ or θ_2/σ_2 or $1-\pi_2$
- If T is positive => stronger effect of SNP1

  If T is negative => stronger effect of SNP2

- *P*-value q_1 for a one-sided test with H_a: $E_2>E_1$ is $p(T)$

  P-value q_2 for a one-sided test with H_a: $E_1>E_2$ is $1-q_1$

- Evidence Ratio: q_1/q_2 in favor of SNP1 compared to SNP2

Evidence Ratio

Evidence Ratio

q_1/q_2 in favor of SNP1 compared to SNP2

Evidence ratio of 95:5 means that we can conclude that $E_1>E_2$ in a one-sided , 5% significance test of H_0: $E_1=E_2$.

Evidence ratio in favor of E_1 vs. E_2

More significant		Log difference between p_1 and p_2						
-log p_1	r	0	1	2	3	4	5	6
10	0.9	68:32	76:24	84:16	89:11	94:6	97:3	99:1
	1.0	50:50	60:40	70:30	79:21	87:13	93:7	97:3
	1.1	33:67	44:56	55:45	67:33	78:22	87:13	94:6
9	0.9	67:33	76:24	84:16	90:10	95:5	98:2	99:1
	1.0	50:50	61:39	71:29	81:19	89:11	95:5	98:2
	1.1	34:66	45:55	57:43	70:30	81:19	90:10	96:4
8	0.9	66:34	76:24	84:16	91:9	95:5	98:2	100:0
	1.0	50:50	61:39	73:27	83:17	91:9	96:4	99:1
	1.1	35:65	47:53	60:40	73:27	85:15	93:7	98:2
7	0.9	65:35	75:25	84:16	92:8	96:4	99:1	100:0
	1.0	50:50	62:38	75:25	85:15	93:7	98:2	100:0
	1.1	36:64	49:51	63:37	77:23	89:11	96:4	99:1
6	0.9	64:36	75:25	85:15	93:7	98:2	100:0	
	1.0	50:50	64:36	77:23	88:12	96:4	99:1	
	1.1	37:63	52:48	67:33	82:18	93:7	99:1	
5	0.9	62:38	75:25	86:14	95:5	99:1		
	1.0	50:50	65:35	80:20	91:9	98:2		
	1.1	39:61	55:45	72:28	87:13	97:3		
4	0.9	61:39	76:24	89:11	97:3			
	1.0	50:50	67:33	84:16	96:4			
	1.1	40:60	59:41	78:22	94:6			
3	0.9	59:41	77:23	93:7				
	1.0	50:50	71:29	90:10				
	1.1	42:58	64:36	87:13				

$$r_1 = \frac{\sigma_2 / \sqrt{n_2}}{\sigma_1 / \sqrt{n_1}}$$

$$r_2 = \sqrt{\frac{n_1}{n_2}}$$

$$r_3 = 1$$