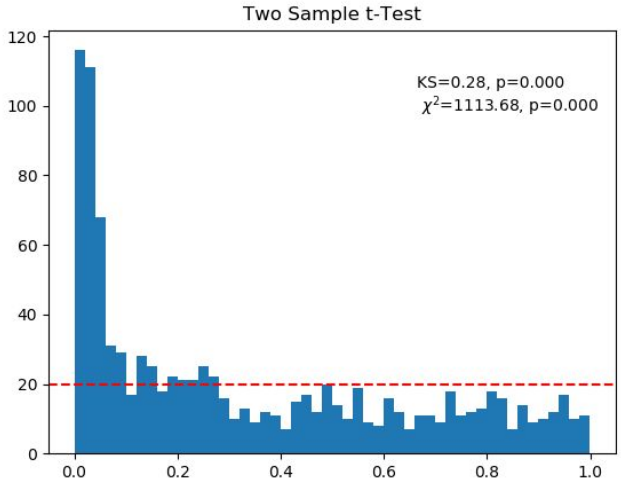
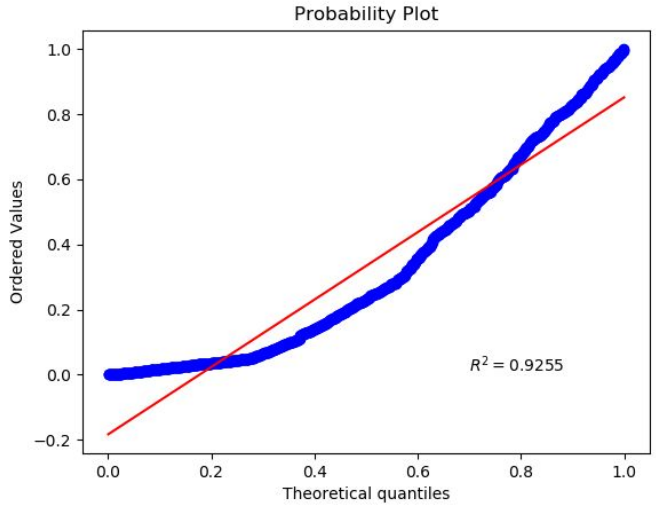
Applying the Delta Method to AB Testing

A Case Study

The Problem: AA Testing Results





AB Testing Review: Two Sample t-Test

Null Hypothesis: $\mu_c = \mu_t$

Alternative Hypothesis: $\mu_c \neq \mu_t$

Two Sample t-Test:

$$T = \frac{\Delta}{\sqrt{Var(\Delta)}}$$

$$\Delta = \mu_t - \mu_c$$

$$var(\Delta) = var(X_t) + var(X_c)$$

AB Testing Review: Variance Estimation

Compute Metric: $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Compute sample variance: $$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Compute variance of avg metric: $$var\left(\overline{X}\right) = \frac{var(X)}{n}$$

Standard Model

Metric
1.2
2.3
2.9
1.9
5.5
5.1
7.4
6.1
8.2
7.8

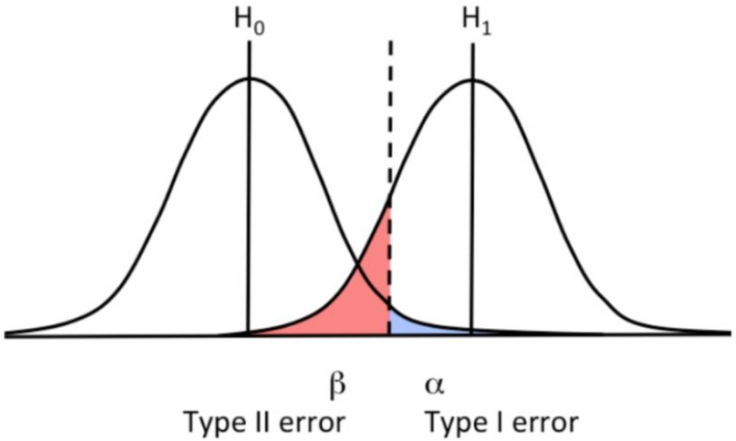
N = 10

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ ⇨ **Mean** = 4.84

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$ ⇨ **Variance** = 6.76

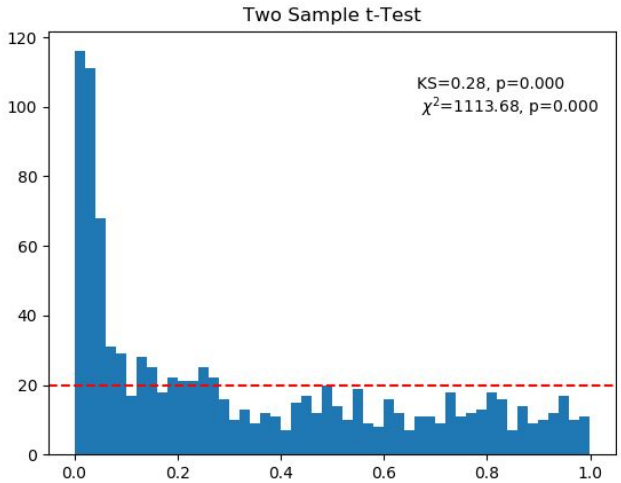
Type I vs. Type II Errors

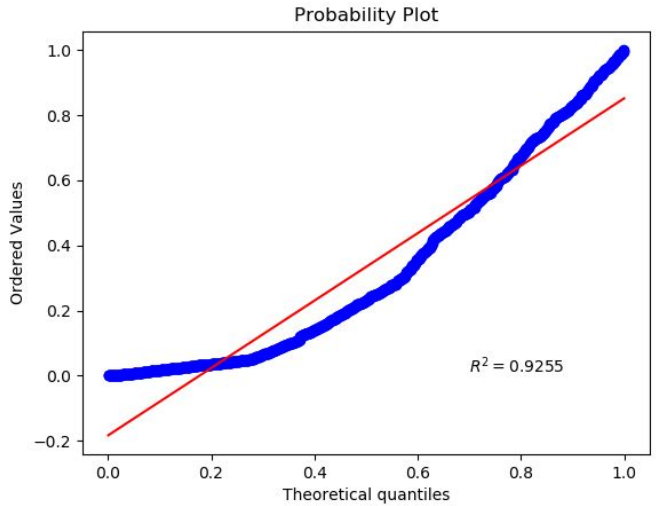
p-value: Probability under H_0 of observing test results as extreme or more extreme than what we have observed



	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

Back to the Problem: AA Testing Results





What's happening? A Theory

The Equation, $Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, is simple to understand, but it has a very important assumption. It must be i.i.d!

This is satisfied when the analysis unit = randomization unit

Often, in online experiments the randomization unit is at the user level, but the analysis unit is at the page or session level.

Example: CTR = Total Clicks / Total page views (analysis unit = page view)

Because the analysis unit ≠ randomization unit, we are systematically underestimating variance

Metric	User
1.2	User 1
2.3	User 2
2.9	User 2
1.9	User 2
5.5	User 3
5.1	User 3
7.4	User 3
6.1	User 3
8.2	User 3
7.8	User 3

N = ?

Mean = ?

Variance = ?

Options we explored

- Aggregate metrics
- Linear Regression (many)
- Delta Method (Microsoft)
- Sequential Model (Walmart)
- Bootstrap methods (many)
- Bayesian Model (many)
- Hierarchical Linear Model
- Method of Moments Hierarchical Linear Model (Stitch Fix)

Delta Method

- Delta method allows us to extend the normal approximations from the central limit theorem broadly.
- “For any random variable T_n (the subscript indicates its dependency on n, e.g., sample average) and constant θ such that $\sqrt{n}(T_n - \theta) \rightarrow N(0, 1)$ in distribution as $n \rightarrow \infty$, the Delta method allows us to extend its asymptotic normality to any continuous transformation $\phi(T_n)$.”
- The only assumption is “big data”
- Benefits
  - Easily programmable
  - Easily parallelizable
  - Highly extensibile

Delta Method: Applied to a Ratio Metric

Write metric as ratio of “average of user level metrics”

$$M = \frac{\overline{X}}{\overline{Y}}$$

Then variance is equal to…

$$var(M) = \frac{1}{\overline{Y}^2} var\left(\overline{X}\right) + \frac{\overline{X}^2}{\overline{Y}^4} var\left(\overline{Y}\right) - 2\frac{\overline{X}}{\overline{Y}^3} cov\left(\overline{X}, \overline{Y}\right)$$

Example: CTR = Total Clicks / Total Pageviews

X = Total of clicks per user

Y = Total page view per user

Delta Method

Metric	User
1.2	User 1
2.3	User 2
2.9	User 2
1.9	User 2
5.5	User 3
5.1	User 3
7.4	User 3
6.1	User 3
8.2	User 3
7.8	User 3

User	Metric sum	User count
1	1.2	1
2	7.1	3
3	40.1	6

N = 10

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ ⇒ **Mean** = 4.84

$$var(M) = \frac{1}{\overline{Y}^2} var(\overline{X}) + \frac{\overline{X}^2}{\overline{Y}^4} var(\overline{Y}) - 2\frac{\overline{X}}{\overline{Y}^3} cov(\overline{X}, \overline{Y})$$ ⇒ **Variance** = 8.58

How to validate new AB testing Method

1. When there is no real effect how often does our model say so?

   AA Testing

2. Do we get reasonable mean and variance estimates?

   Simulation (of distributions with known means and variances)

3. When there is a real effect how quickly does our model say so?

   Power Analysis (comparison of distributions with known mean effects)

Conclusion

- The number of False Positives in AA Experiments went from 28% to 6%
- Even simple models can become complicated when applied in the real world
- Lots of great resources available; Don’t try to reinvent the wheel
- Most sophisticated model not always best
- Performance matters a lot in real world settings (duh)
- A huge thanks to Ron Kohavi for letting me preview his as of yet unreleased book, “Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing”. Much of the material here was inspired by this book