Hypothesis Testing
Part I

James J. Heckman
University of Chicago

Econ 312, Spring 2021

1. A Brief Review of Hypothesis Testing and Its Uses
Common Phrase:
Chicago Economics test Models
What are Valid Tests?

- Key Distinction: $\underbrace{\text{ex ante}}_{\substack{\text{classical}\\ \text{inference}}}$ vs. $\underbrace{\text{ex post}}_{\substack{\text{likelihood}\\ \text{principle;}\\ \text{Bayesian}\\ \text{inference}}}$ inference

- P values and pure significance tests (R.A. Fisher)—focus on null hypothesis testing.
- Neyman-Pearson tests—focus on null and alternative hypothesis testing.
- Both involve an appeal to long run trials. They adopt an *ex ante* position (justify a procedure by the number of times it is successful if used repeatedly).

2. Pure Significance Tests

- Focuses exclusively on the null hypothesis
- Let $(Y_1, \ldots, Y_N)$ be observations from a sample.
- Let $t(Y_1, \ldots, Y_N)$ be a test statistic.
- If
  1. We know the distribution of $t(\underset{\sim}{Y})$ under H_0, and
  2. The larger the value of $t(\underset{\sim}{Y})$, the more the evidence against H_0,
- Then

$$P_{obs} = \Pr(T \geq t_{obs} : H_0).$$

- Then a high value of P_{obs} is evidence against the null hypothesis.
  - Observe that under the null P value is a uniform $(0, 1)$ variable.
  - For random variable with density (absolutely continuous with Lebesgue measure) $Z = F_X(X)$ is uniform for any X given that F_X is continuous.
  - *Prove this. It is automatic from the definition.*
  - P value — probability that T would occur given that H_0 is a true state of affairs.
  - F test or t test for a regression coefficient is an example.

- - The higher the test statistic, the more likely we reject.
  - Ignores any evidence on alternatives.
  - R.A. Fisher liked this feature because it did not involve speculation about other possibilities than the one realized.
  - P values make an absolute statement about a model.
- *Questions to consider*:
  1. How to construct a 'best' test? Compare alternative tests. Any monotonic transformation of the "t" statistic produces the same P value.
  2. Pure significance tests depend on the sampling rule used to collect the data. This is not necessarily bad.
  3. How to pool across studies (or across coefficients)?

2.1 Bayesian vs. Frequentist vs. Classical Approach

- ISSUES:
    1. In what sense and how well do significance levels or "P" values summarize evidence in favor of or against hypotheses?
    2. Do we always reject a null in a big enough sample? Meaningful hypothesis testing—Bayesian or Classical—requires that "significance levels" decrease with sample size;
    3. Two views: $\beta = 0$ tests something meaningful vs. $\beta = 0$ only an approximation, shouldn't be taken too seriously.
    4. $Y = X\beta + U$

4. How to quantify evidence about model? (How to incorporate prior restrictions?) What is “strength of evidence?”
5. How to account for model uncertainty: “fishing,” etc.

- First consider the basic Neyman-Pearson structure- then switch over to a Bayesian paradigm.

- Useful to separate out:
    1. Decision problems.
    2. Acts of data description.
- This is a topic of great controversy in statistics.

- *Question:* In what sense does increasing sample size always lead to rejection of an hypothesis?
    - If null not exactly true, we get rejections (The power of test $\rightarrow 1$ for fixed sig. level as sample size increases)
- Example to refresh your memory about Neyman-Pearson Theory.
- Take one-tail normal test about a mean:
- What is the test?

$$\begin{aligned} H_0 &: \bar{X} \sim N\left(\mu_0, \sigma^2/T\right) \\ H_A &: \bar{X} \sim N\left(\mu_A, \sigma^2/T\right) \end{aligned}$$

- Assume σ^2 is known.

- For any c we get

$$\Pr\left(\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/T}} > \frac{c-\mu_0}{\sqrt{\sigma^2/T}}\right) = \alpha(c).$$

- (We exploit symmetry of standard normal around the origin).
- For a fixed α, we can solve for $c\left(\alpha\right)$.

$$c\left(\alpha\right) = \mu_0 - \frac{\sigma}{\sqrt{T}}\Phi^{-1}\left(\alpha\right).$$

- Now what is the probability of rejecting the hypothesis under alternatives? (The power of a test).
- Let μ_A be the alternative value of μ_A.
- Fix c to have a certain size. (Use the previous calculations)

$$\Pr\left(\frac{\bar{X}-\mu_A}{\sqrt{\sigma^2/T}} > \frac{c-\mu_A}{\sqrt{\sigma^2/T}}\right) = \Pr\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A-\frac{\sigma}{\sqrt{T}}\Phi^{-1}(\alpha)}{\left(\sigma/\sqrt{T}\right)}\right).$$

- We are evaluating the probability of rejection when we allow μ_A to vary.

- Thus

$$\begin{aligned} &= \Pr\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A}{\left(\sigma/\sqrt{T}\right)} - \Phi^{-1}(\alpha)\right) \\ &= \alpha \qquad \text{when } \mu_0 = \mu_A \end{aligned}$$

- If $\mu_A > \mu_0$, this probability goes to one.
- This is a *consistent test*.

- Now, suppose we seek to test $H_0 : \mu_0 > k$.
- Use

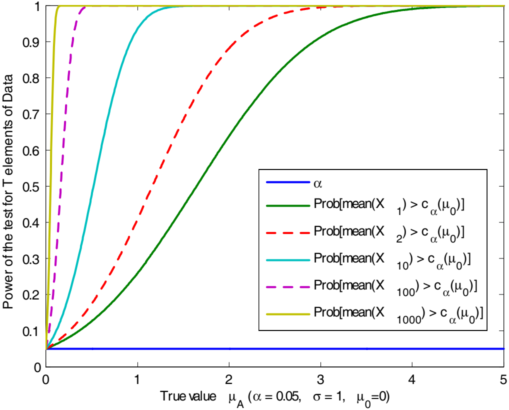
$$\bar{X} > k, \text{ fixed } k$$

- If μ_0 is true:

$$\frac{\bar{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{T}}\right)} > \frac{k - \mu_0}{\left(\frac{\sigma}{\sqrt{T}}\right)}$$

- The distribution becomes more and more concentrated at μ_0.
- We reject the null unless $\mu_0 = k$.

Probability of Rejecting H_0



$$\Pr\left(\overline{X}_T > c_\alpha\left(\mu_0\right)\right)$$

$$\overline{X}_T \sim N(\mu_A, \frac{\sigma^2}{T}), \quad c_\alpha\left(\mu_0\right) = \mu_0 - \frac{\sigma}{\sqrt{T}}\Phi^{-1}\left(\alpha\right)$$

- Parenthetical Note:
- Observe that if we measure X with the slightest error and the errors do not have mean zero, we always reject H_0 for T big enough.

Design of Sample size

- Suppose that we fix the power $= \beta$.
- Pick $c(\alpha)$.
- What sample size produces the desired power?
- We postulate the alternative $= \mu_0 + \Delta$.

$$\Pr\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A}{\left(\sigma/\sqrt{T}\right)} - \Phi^{-1}(\alpha)\right)$$

$$= \Phi\left(\Phi^{-1}(\alpha) + \frac{\mu_A-\mu_0}{\frac{\sigma}{\sqrt{T}}}\right) = \beta$$

$$\begin{aligned} \Phi^{-1}(\beta) &= \Phi^{-1}(\alpha) + \frac{\mu_A-\mu_0}{\frac{\sigma}{\sqrt{T}}} \\ \frac{\left[\Phi^{-1}(\beta) - \Phi^{-1}(\alpha)\right]}{\left(\frac{\Delta}{\sigma}\right)} &= \sqrt{T} \end{aligned}$$

- Minimum T needed to reject null at specified alternative.
- Has power of β for "effect" size Δ/σ.
- Pick sample size on this basis: (This is used in sample design)
- What value of β to use?
- Observe that two investigators with same α but different sample size T have different power.
- This is often ignored in empirical work.
- Why not equalize the power of the tests across samples?
- Why use the same size of test in all empirical work?

3. Alternative Approaches to Testing and Inference

3.1 Classical Hypothesis Testing

1. Appeals to *long run frequencies*.
2. Designs an *ex ante* rule that *on average* works well. e.g. 5% of the time in repeated trials we make an error of rejecting the null for a 5% significance level.
3. Entails a hypothetical set of trials, and is based on a long run justification.

(4) Consistency of an estimator is an example of this mindset. E.g., $Y = X\beta + U$

$$E\left(U \mid X\right) \neq 0\text{; OLS biased for } \beta.$$

Suppose we have an instrument:

$$Cov\left(Z, U\right) = 0 \qquad Cov\left(Z, X\right) \neq 0$$

$$\text{plim } \beta_{OLS} = \beta + \frac{Cov\left(X, U\right)}{Var\left(X\right)}$$

$$\text{plim } \beta_{IV} = \beta + \underbrace{\frac{Cov\left(Z, U\right)}{Cov\left(Z, X\right)}}_{=0} = \beta$$

- Because $Cov\left(Z, U\right) = 0$.
- Assuming $Cov\left(Z, X\right) \neq 0$.

- Another consistent estimator
    1. Use *OLS* for first 10^{100} observations
    2. Then use *IV*.
- Likely to have poor small sample properties.
- But on a long run frequency justification, its just fine.

3.2 Examples of why some people get very unhappy about classical testing procedures

Classical inference: ex ante

Likelihood and Bayesian statistics: ex post

Example 1.

(Sample size: $T = 2$)

$$(X_1, X_2) \qquad X_1 \perp\!\!\!\perp X_2.$$

$$P_{\theta_0}\left(X = \theta_0 - 1\right) = P_{\theta}\left(X = \theta_0 + 1\right) = \frac{1}{2}$$

- One possible (smallest) confidence set for θ_0 is

$$C\left(X_1, X_2\right) = \begin{cases} \frac{1}{2}\left(X_1 + X_2\right) & \text{if} \quad X_1 \neq X_2 \\ X_1 - 1 & \text{if} \quad X_1 = X_2 \end{cases}$$

- Thus 75% of the time $C\left(X_1, X_2\right)$ contains θ_0 (75% of repeated trials it covers θ_0). (Verify this)
- Yet if $X_1 \neq X_2$, we are *certain* that the confidence interval exactly covers the true value 100% of the time it is right.
- *Ex post* or conditional inference on the data, we get the exact value.

Example 2.

(D.R. Cox)

1. You have data, say on DNA from crime scenes.
2. You can send data to New York or California labs. Both labs seem equally good.
3. Toss a coin to decide which lab analyzes data.

- Should the coin flip be accounted for in the design of the test statistic?

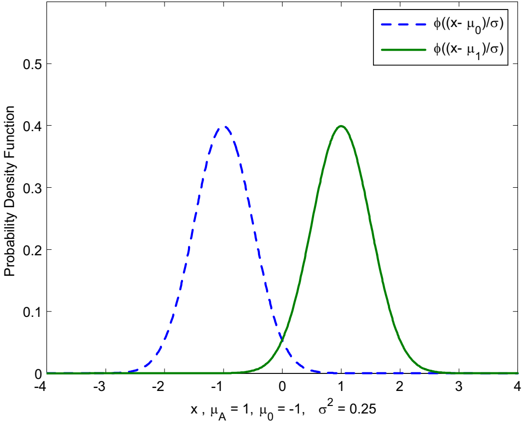
Example 3.

- Test $H_0 : \theta = -1$ vs. $H_A : \theta = 1$; $X \sim N(\theta, .25)$
- Consider rejection region: Reject if $X \geq 0$.
- If we observe $X = 0$, we would have $\alpha = .0228$
- Size is .0228; power under alternative is .9772. Consistent test; unbiased test. Looks good.
- If we reverse roles of null and alternative, it would also look good.

Example 3

$$X \sim N(\mu, 0.25); \quad H_0 : \mu = -1; H_A : \mu = 1.$$

$$\textit{Test} : \text{Reject } H_0 \text{ if } X \geqslant 0.$$



In the case of 0 being observed: Power $= \alpha = 0.0228$

Example 4.

Likelihood Principle vs. Classical Inference

- $X \in \{1, 2, 3\}$; we have two possible models (nulls and alternatives): "0" and "1."

	1	2	3
P_0	.009	.001	.99
P_1	.001	.989	.01

← values random variables can assume

- Consider the following test:
- Accepts P_0 when $X = 3$ and accepts P_1 otherwise
- ($\alpha = .01$ and $\beta = .99$ high power).
- Unbiased and consistent test.

- If we observe $X = 1$ we reject H_0.
- But the likelihood ratio in favor of "0" is

$$\frac{.009}{.001} = 9$$

- **Likelihood principle:** alternative inferential criterion.
- **All of the sample information is in likelihood.**

Example 5.

(Likelihood Principle)

	1	2	3
P_0	.005	.005	.99
P_1	.0051	.9489	.01

- Reject "0" when $X = 1, 2$
- Power $=$ .99, Size $=$ .01.
- Is it reasonable to pick "1" over "0" when $X = 1$ is observed? (Likelihood ratio not strongly supporting the hypothesis)

Example 6.

(Lindley and Phllips; American Statistician, August, 1976): Irrelevance of Stopping Rules in the Likelihood Principle and in Bayesian Analysis.

- Consider an experiment.
- We draw 12 balls from an urn. The urn has an infinite number of balls.
- $\theta =$ probability of black.
- $(1 - \theta) =$ probability of red.

- Null hypothesis: Red and black are equally likely on each trial and trials are independent.

$$\Pr\left(X \text{ is black}\right) = \binom{12}{X} \theta^{X} \left(1-\theta\right)^{12-X}.$$

- Suppose that we draw 9 black balls and 3 red balls.
- What is the evidence in support of the hypothesis that $\theta = \frac{1}{2}$?

- Consider a critical region $X = \{9, 10, 11, 12\}$ to reject null of $\theta = \frac{1}{2}$: $(H_0; \theta = \frac{1}{2})$

$$\begin{aligned} \alpha &= \Pr\left(X \in \{9, 10, 11, 12\}\right) \\ &= \left\{ \binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0} \right\} \left(\frac{1}{2}\right)^{12} \doteq 7.5\% \end{aligned}$$

- We do not reject H_0 using $\alpha = .05$.
- Would reject if we chose $\alpha = .10$
- This sampling distribution assumes that 10 black and 2 reds is a possibility.
- It is based on a counterfactual space of what else could occur and with what possibility.

- Consider an alternative sampling rule.
- Draw balls until 3 red balls are observed and then stop.
- So 10 blacks and 2 reds on a trial of 12 observations not possible as they were before.
- Distribution of X_2 (X in this experiment) is

$$\binom{X_2+2}{X_2}\theta^{X_2}(1-\theta)^3$$

- Prove this (negative binomial).

- Use same rejection region $X_2 = \{9, 10, 11, 12, 13, \dots\}$ i.e., if $X_2 \geq 9$, reject
- Note:
$$\Pr(X \in \{9, 10, 11, 12, 13, ...\}) = 3.25\%$$
- Now "significant." Reject null of $\theta = \frac{1}{2}$.
- In both cases 9 black and 3 red on a single trial.
- They are the same for a Bayesian (will show below).
- They have the same m/e independent of stopping rule

- **In computing** P **values and significance levels, you need to model what didn't occur**.
- Depends on the stopping rule and the hypothetical admissible sample space.

3.3 Likelihood Principle

- All of the information is in the sample.
- Look at the likelihood as best summary of the sample.

Likelihood Approach

- Recall from your previous lectures of asymptotics that under the regularity conditions $Q_T(\theta)$ is a valid criterion:

$$\begin{aligned} Q_T(\hat{\theta}) &= Q(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' \left. \frac{\partial^2 Q_T}{\partial\theta\partial\theta'} \right|_{\theta_0} (\hat{\theta} - \theta_0) + o_P(1) \\ \text{because } \frac{\partial Q_T}{\partial \hat{\theta}} &= 0 \text{ for all } \hat{\theta}; \\ & \text{For likelihood } \mathcal{L}: \\ Q_T &= \frac{\ln \mathcal{L}(\hat{\theta})}{T} \\ Q(\theta_0) &= \frac{\ln \mathcal{L}(\theta_0)}{T} \end{aligned}$$

- In terms of the information matrix, for the likelihood case

$$Q_T(\hat{\theta}) = Q(\theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)' I_{\theta_0}(\hat{\theta} - \theta_0) + o_P(1)$$

- So we know that as $T \to \infty$, the normalized likelihood $\mathcal{L}$ converges to a normal, e.g.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^k \left|\Sigma\right|^{\frac{k}{2}}} \exp\left[-\frac{1}{2}(X - \mu)'\Sigma^{-1}(X - \mu)\right]$$

- $$\ln \mathcal{N}(\mu, \Sigma) = -k \ln(2\pi) - \frac{k}{2} \ln |\Sigma| - \frac{1}{2}(X - \mu)'\Sigma^{-1}(X - \mu).$$
- So the likelihood is converging to a normal-looking criterion and has its mode at θ_0.
- The most likely value is at the *MLE* estimator (mode of likelihood is θ_0).
- In the example of 9 black and 3 red, we have same $\hat{\theta}_0$ for either stopping rule: likelihood ignores constants.

Bayesian Principle

- Use prior information in conjunction with sample information.
- Place priors on parameters.
- Classical Method and Likelihood Principle sharply separate parameters from data (random variables).
- The Bayesian method does not.
- All parameters are random variables.

- Bayesian and Likelihood approach both use likelihood.
- Likelihood: Use data from experiment.
- Evidence concentrates on θ_0.
- For both Bayesians and likelihood principle inference: irrelevance of stopping rules.
- Bayesian: Use data from experiment plus prior.
- Bayesian Approach postulates a prior $p(\theta)$.
- This is a probability density of θ.

- Compute using posterior (Bayes Theorem):

$$\overbrace{\pi\left(\theta \mid X\right)}^{\text{posterior}} = \mathcal{T}\,\underbrace{\mathcal{L}\left(\theta \mid X\right)}_{\text{likelihood}}\overbrace{p\left(\theta\right)}^{\text{prior}}$$

- Where $\mathcal{T}$ is a constant defined so posterior integrates to 1.
- Get some posterior independent of constants (and therefore sampling rule).

Definetti's Thm:

- Let X_i denote a binary variable $X_i \in \{0, 1\}$, X_i i.i.d.
- $\Pr(X_i = 1) = \theta$
- $\Pr(X_i = 0) = 1 - \theta$
- Let $p(r, s) =$ probability of r "1s" and s "0s": total number of balls $(r + s)$ drawn.
- If series is *exchangeable*

$$p(r, s) = \int_0^1 \binom{r+s}{r} \theta^r (1 - \theta)^s p(\theta) d\theta$$

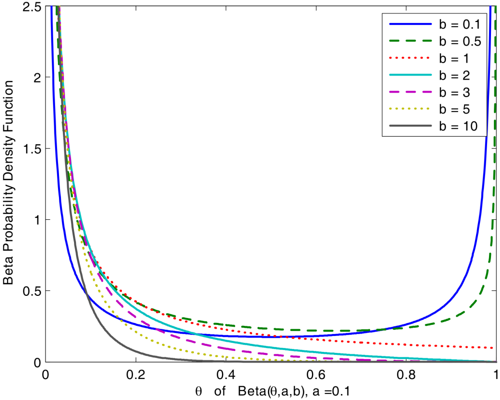
- Therefore, there exists a heterogeneity distribution.
- For some $p(\theta) \geq 0$ (this is just the standard Hausdorff moment problem).

Conjugate Priors

- For this problem a natural "conjugate" prior is

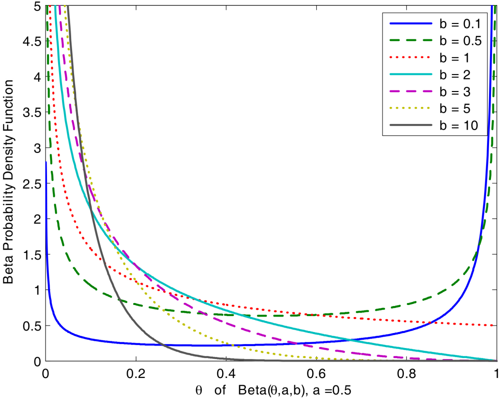
$$\begin{aligned} p(\theta) &= \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \qquad\qquad 0 \leq \theta \leq 1 \\ a &= b = 1, \text{ uniform} \\ E(\theta) &= \frac{a}{a+b} \end{aligned}$$

The Beta Probability Density Function
Beta 1



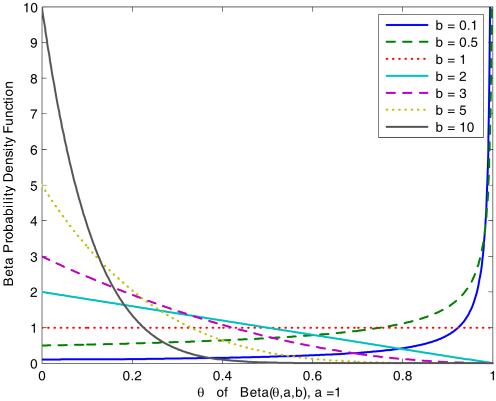
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}; a = 0.1;$$

The Beta Probability Density Function
Beta 2



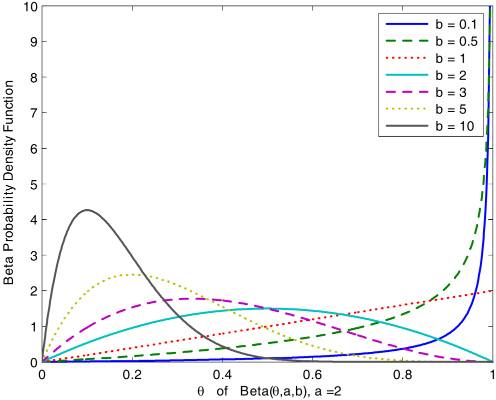
$$BetaPDF(\theta,a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a=0.5;$$

The Beta Probability Density Function
Beta 3



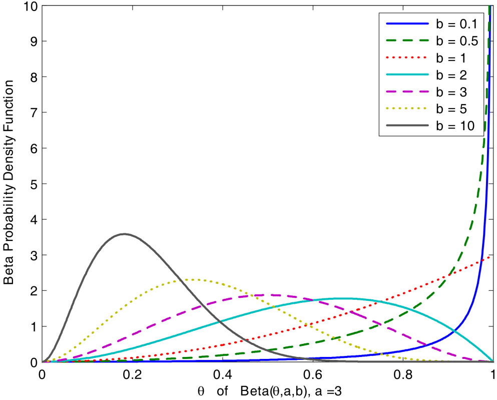
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 1;$$

The Beta Probability Density Function
Beta 4



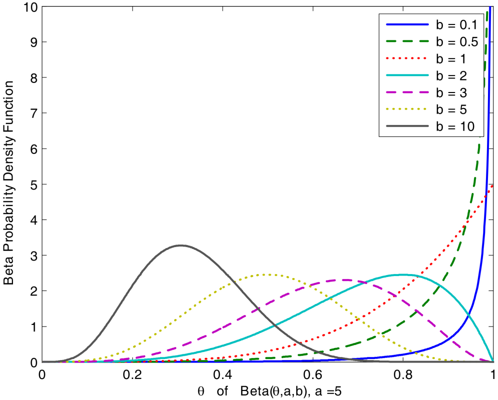
$$BetaPDF(\theta,a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a=2;$$

The Beta Probability Density Function
Beta 5



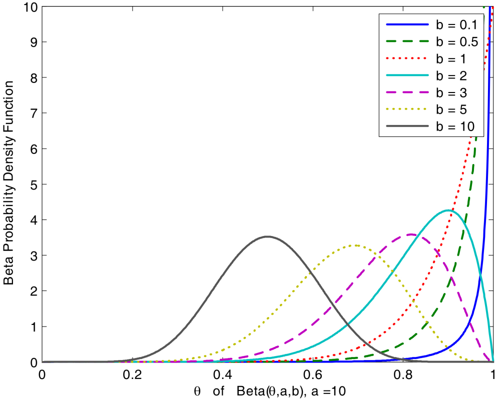
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 3;$$

The Beta Probability Density Function
Beta 6



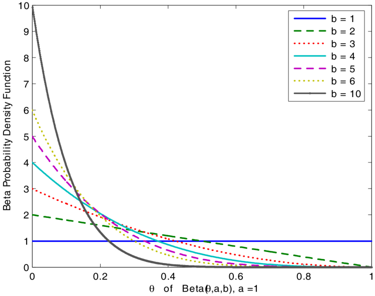
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 5;$$

The Beta Probability Density Function
Beta 7



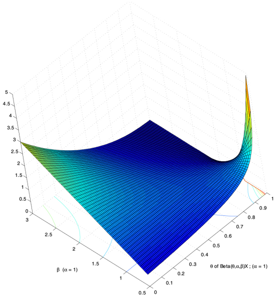
$$BetaPDF(\theta,a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 10;$$

The Beta Probability Density Function
Beta 8



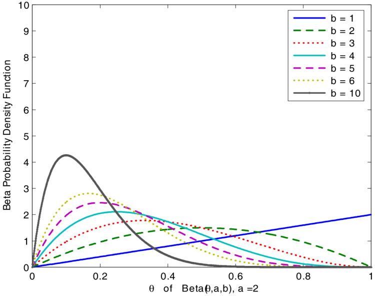
$$BetaPDF(\theta,a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\;\;(a=1)\,;$$

The Beta Probability Density Function
Beta 9



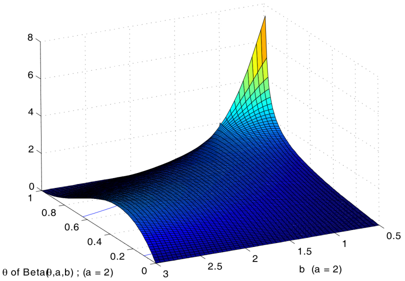
$$BetaPDF(\theta,a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\;b\in[0.5,3]\,;$$

The Beta Probability Density Function
Beta 10



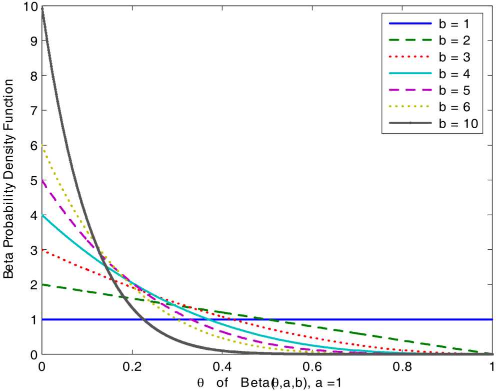
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 2;$$

The Beta Probability Density Function
Beta 11



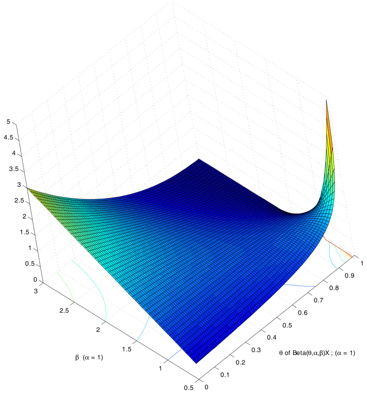
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ b \in [0.5, 3]\,;$$

The Beta Probability Density Function
Beta 8



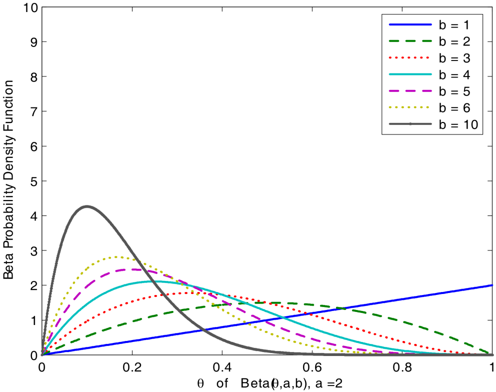
$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}; \quad (a = 1);$$

The Beta Probability Density Function
Beta 9



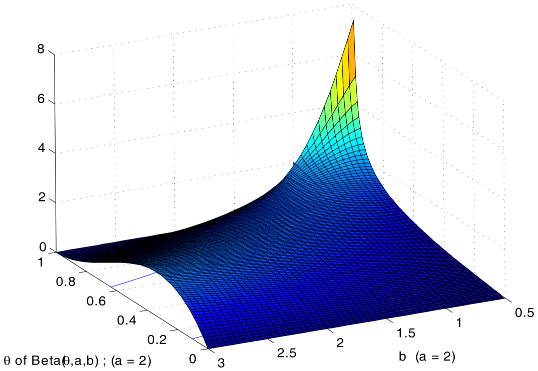
$$BetaPDF(\theta,a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ b \in [0.5,3];$$

The Beta Probability Density Function
Beta 10



$$BetaPDF(\theta, a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ a = 2;$$

The Beta Probability Density Function
Beta 11



$$BetaPDF(\theta,a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)};\ b\in[0.5,3];$$

Bayesian Posterior Density

- Posterior

$$\pi(\theta \mid X) = \tau \underbrace{\theta^{r}(1-\theta)^{s}}_{\text{likelihood}} \underbrace{\theta^{a-1}(1-\theta)^{b-1}}_{\text{prior}},$$

where X is the data and τ is a normalizing constant to make density normalize to one:

$$\tau \int \theta^{r}(1-\theta)^{s}\, \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 1$$

- Observe crucially that the normalizing constant is the same for both sampling rules we discussed in the red ball and black ball problem.

- Why? Because we choose τ to make $\pi(\theta \mid X)$ integrate to one.
- Mean of posterior with prior a, b

$$E^{\text{ posterior}}(\theta) = \frac{a+r}{(a+r)+(b+s)}$$

- **Notice: The constants that played such a crucial role in the sampling distribution play no role here. They vanish in defining the constant τ.**

$$\text{mode of } \theta = \frac{a+r-1}{(a+r-1)+(b+s-1)}$$

- Likelihood corresponds to $(r+s)$ trials with r red and s black.
- Prior corresponds to $(a+b-2)$ trials with $(a-1)$ red and $(b-1)$ black.

Empirical Bayes Approach

- Estimate "Prior".
- Go to Beta-Binomial Example.

$$p(r,s)=\int_0^1 \frac{\binom{r+s}{r}\theta^r(1-\theta)^s\theta^{a-1}(1-\theta)^{b-1}}{B(a+b)}d\theta.$$

- Now θ is a heterogeneity parameter distributed $B\left(a,b\right)$.

$$=\frac{\binom{r+s}{r}B(a+r-1,b+s-1)}{B(a+b)}$$

- Estimate a and b as parameters from a string of trials with r reds and s blacks. θ is a person-specific parameter.

- Similar idea in the linear regression model $Y_i = X_i\beta_i + \varepsilon_i$.

Random Coefficient Regression

- We can identify means and variances of β.

$$Y_i = X_i \beta_i + \varepsilon_i \qquad X_i \perp\!\!\!\perp (\beta_i, \varepsilon_i)$$

$$\beta_i = \bar{\beta} + U_i \qquad E(U_{(i)} U'_{(i)}) = \Sigma_U$$

- Assume $\varepsilon_i \perp\!\!\!\perp \beta_i$; $X_i \perp\!\!\!\perp \varepsilon_i$.

$$Y_i = X_i \bar{\beta} + \underbrace{(X_i U_i + \varepsilon_i)}_{\nu_i}$$

$$E\left[\nu_i^2 \mid X_i\right] = \sigma_\varepsilon^2 + X_i \Sigma_U X_i'$$

- Use squared OLS residuals to identify Σ_U given X.

- Notice: We can extend the model to allow

$$\beta_i = \Phi Z_i + U_i$$

and identify Φ (Hierarchical model).

- **Digression:** Take the Classical Normal Linear Regression Model

$$Y = X\beta + U, \qquad U \perp\!\!\!\perp X, \qquad E(UU') = \sigma^2 I$$

$$OLS \quad \hat{\beta} = (X'X)^{-1}X'Y \qquad Var(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

- Assume σ^2 known. Take a conjugate prior on β.

$$\beta \sim \mathcal{N}\left(\bar{\beta}, \sigma^2(C)^{-1}\right)$$

- Posterior is normal:

$$\beta_{\text{posterior}} \sim \mathcal{N}\left(\left(C + (X'X)^{-1}\right)^{-1}\left(C\bar{\beta} + (X'X)\hat{\beta}\right), \sigma^2(C + X'X)^{-1}\right)$$

- Thus, we can think of the prior as a sample of observations with the "$(X'X)$" matrix being C and the "sample" OLS from prior being $\bar{\beta}$.

- Compare to

$$\begin{bmatrix} Y^* \\ Y \end{bmatrix} = \begin{bmatrix} X^* \\ X \end{bmatrix} \beta + \begin{bmatrix} U^* \\ U \end{bmatrix}.$$

- OLS is $\quad (X^{*\prime}X^* + X'X)^{-1} (X^{*\prime}X^* b^* + X'Xb)\,,$

$$b^* = (X^{*\prime}X^*)^{-1} X^{*\prime}Y^*, \qquad b = (X'X)^{-1} X'Y.$$

- (Prove this.)
- In other words see, e.g., Robert (Bayesian choice) for more general case where σ^2 is unknown (gamma prior).

- To compute evidence on one hypothesis vs. another hypothesis use posterior odds ratio

$$\frac{\Pr(H_1 \mid X)}{\Pr(H_0 \mid X)} = \frac{\Pr(X \mid H_1)}{\Pr(X \mid H_0)} \frac{\Pr(H_1)}{\Pr(H_0)}$$

- Hypotheses are restrictions on the prior (e.g. different values of (a, b))

	Classical Approach	Bayesian Approach
Assumption regarding experiment	Events independent, given a probability	Events form exchangeable sequences
Interpretation of probability	Relative frequency; applies only to repeated events	Degrees of belief; applies both to unique and to sequences of events
Statistical inferences	Based on sampling distribution; sample space or stopping rule must be specified	Based on posterior distribution; prior distribution must be assessed
Estimates of parameters	Requires theory of estimation	Descriptive statistics of the posterior distribution
Intuitive judgement	Used in setting significance levels, in choice of procedure, and in other ways	Formally incorporated in the prior distribution

Source: Lindley, D.V. and Phillips, L.D. (1976). "Inference for a Bernoulli Process (A Bayesian View)." *American Statistician* 30(3): 112-119

Bayesian Testing
Point null vs. Point Alternative test

- Think of a regression model $Y = X\beta_1 + U_1$ vs. $Y = X\beta_0 + U_0$
- 2 Hypotheses: H_1, H_0

$$\overset{\text{Posterior odds ratio}}{\frac{\Pr(H_1 \mid Y)}{\Pr(H_0 \mid Y)}} = \overset{\text{Bayes factor}}{\frac{\Pr(Y \mid H_1)}{\Pr(Y \mid H_0)}} \overset{\text{Prior odds ratio}}{\frac{\Pr(H_1)}{\Pr(H_0)}}$$

- "Predictive density":

$$f\left(Y \mid H_i\right) = \int_{\underset{\sim}{\beta}_i} \int_{\sigma_i^2} \underset{\text{Likelihood}}{f\left(Y \mid H_i, \beta_i, \sigma_i^2\right)} \, \underset{\text{Prior density}}{f\left(\beta_i, \sigma_i^2\right)} \, d\beta_i \, d\sigma_i$$

- Evidence supports the higher posterior probability model.
- Example:

$$Y_i \sim N\left(\mu; \sigma^2\right) \qquad \bar{Y} \sim N\left(\mu; \sigma^2/T\right)$$

$$\begin{aligned}
H_0 &: \mu_0 = 0, \sigma = 1 \\
H_1 &: \mu_1 = 1, \sigma = 1 \\
H_0 &: \bar{Y} \sim N(0, 1/T) \\
H_1 &: \bar{Y} \sim N(1, 1/T)
\end{aligned}$$

- Typical Neyman-Pearson Rule:

$$\begin{aligned} \text{Reject } H_0 \text{ if } \bar{Y} &\geq c \\ \text{Accept } H_0 \text{ if } \bar{Y} &< c \end{aligned}$$

- Type 1 and Type 2 errors:

$$\begin{aligned} \alpha(c) &= \Pr\left(\bar{Y} > c \mid \mu = 0\right) \\ \beta(c) &= \Pr\left(\bar{Y} \leq c \mid \mu = 1\right) \end{aligned}$$

- Example: $c = 0.5$, $\alpha = \beta = 0.31$ (show this).

Bayes Approach

$$\begin{aligned}
\Pr\left(H_0 \mid \bar{Y}\right) &= \frac{f\left(\bar{Y} \mid H_0\right)\Pr\left(H_0\right)}{f\left(\bar{Y}\right)} \\
&= \frac{f\left(\bar{Y} \mid H_0\right)\Pr\left(H_0\right)}{f\left(\bar{Y} \mid H_0\right)\Pr\left(H_0\right) + f\left(\bar{Y} \mid H_1\right)\Pr\left(H_1\right)}
\end{aligned}$$

$$\Pr\left(H_1 \mid \bar{Y}\right) = \frac{f\left(\bar{Y} \mid H_1\right)\Pr\left(H_1\right)}{f\left(\bar{Y} \mid H_0\right)\Pr\left(H_0\right) + f\left(\bar{Y} \mid H_1\right)\Pr\left(H_1\right)}$$

$$
\begin{aligned}
\frac{\Pr\left(H_0 \mid \bar{Y}\right)}{\Pr\left(H_1 \mid \bar{Y}\right)} &= \frac{f\left(\bar{Y} \mid H_0\right)\Pr\left(H_0\right)}{f\left(\bar{Y} \mid H_1\right)\Pr\left(H_1\right)} \\
&= \exp\frac{1}{2}\left[-T\left(\bar{Y}\right)^2 + T\left(\bar{Y}-1\right)^2\right]\left[\frac{\Pr\left(H_0\right)}{\Pr\left(H_1\right)}\right] \\
&= \exp\frac{1}{2}\left[T\bar{Y}^2 - 2T\bar{Y} + T - T\bar{Y}^2\right]\left[\frac{\Pr\left(H_0\right)}{\Pr\left(H_1\right)}\right] \\
&= \left[\exp\frac{1}{2}\left(T - 2T\bar{Y}\right)\right]\left[\frac{\Pr\left(H_0\right)}{\Pr\left(H_1\right)}\right]
\end{aligned}
$$

- Recall $\sigma^2 = 1$ under null and alternatives.

$$\begin{aligned}
\ln\left(\frac{\Pr\left(H_0 \mid \bar{Y}\right)}{\Pr\left(H_1 \mid \bar{Y}\right)}\right) &= \ln\left(\frac{\Pr\left(H_0\right)}{\Pr\left(H_1\right)}\right) + \frac{T}{2}\left(1 - 2\bar{Y}\right) \\
\frac{T}{2}\left(1 - 2\bar{Y}\right) + \ln\left(\frac{\Pr\left(H_0\right)}{\Pr\left(H_1\right)}\right) &> 0 \text{ (If true accept } H_0\text{)} \\
\frac{1}{2} + \frac{\left[\ln\left(\frac{\Pr(H_0)}{\Pr(H_1)}\right)\right]}{T} &> \bar{Y}
\end{aligned}$$

- As T gets big cut off changes with sample size unless $\Pr(H_0) = \Pr(H_1) = \frac{1}{2}$
- Notice that this is different from the classical statistical rule of a fixed cutoff point.

Point Null vs. Composite Alternative

- Same set up as in previous case: $\bar{Y} \sim N(\mu, \sigma^2/T)$.
- $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$. σ^2 is unspecified, but common across models.

- Turn Bayes Crank. Likelihood factor:

$$\frac{f_T\left(Y;\mu,\sigma^2 I\right)}{f_T\left(Y;0,\sigma^2 I\right)}$$

- Relative likelihoods

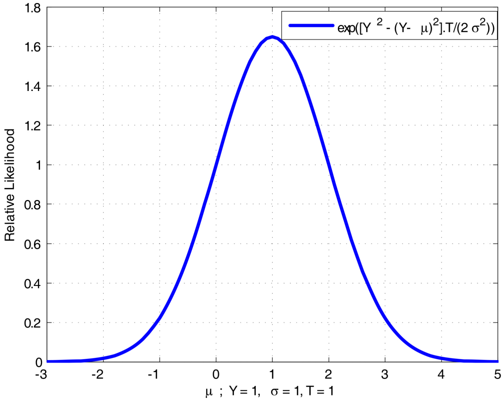
$$\mathcal{L}_R = \exp\left[\frac{T}{2\sigma^2}\left[\bar{Y}^2 - \left(\bar{Y} - \mu\right)^2\right]\right]$$

- What value of μ is best supported by data?
- Recall the likelihood approach: (Focuses on outcomes that are most likely.)

$$\mathcal{L}_R = \exp\left[\frac{T}{2\sigma^2}\mu(2\bar{Y} - \mu)\right]$$

Relative Likelihood for the Model

$$\pounds = \exp(\frac{T}{2\sigma^2}[\overline{Y}^2 - \left(\overline{Y} - \mu\right)^2])$$



- P value approach uses absolute likelihood – not relative likelihood.
- In what sense is it most likely? Likelihood approach:
- Evaluate at null of $\mu = 0$ and we get:

$$\mathcal{L} = \exp\left[-\frac{T}{2\sigma^2}\bar{Y}^2\right] \doteq 1 - \frac{T}{2\sigma^2}\bar{Y}^2 = 1 - \frac{1}{2}\underbrace{\left(\frac{\bar{Y}}{\sqrt{\frac{\sigma^2}{T}}}\right)^2}_{t^2 \text{ for } \mu=0},$$

- This is an expression of support for the hypothesis: $\mu = 0$.
- Thus a big "t" value leads to rejection of the null.
- But this approach does not worry about the alternative.

Frequency Theory or Sampling Approach.

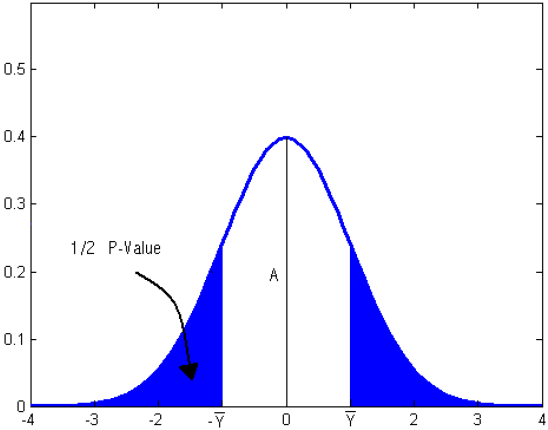
- Look at sampling distributions of model
- Test statistic $\bar{Y}$: centered at $\mu = 0$

$$\alpha(c) = \Pr\left(\bar{Y} > c \mid \mu = 0\right)$$

e.g. $\bar{Y} \geq 1.96 \frac{\sigma}{\sqrt{T}}$ we reject.

- p value: knife-edge value is the value that occurred—value that favors null? At any level less than c, null hypothesis is not rejected.

Sampling Distribution of $\overline{Y}$ (Two sided Test)



- Significance level: is what occurred unlikely?
- Relative likelihood computes evidence of one hypothesis relative to another (null vs. alternative).
- Support for one hypothesis vs. support for another.
- Suppose we allocate positive probability to null.

- Otherwise the probability of a point null $= 0$.

$$P(\mu) \left\{ \begin{array}{ll} \pi & \text{if} \quad \mu = 0 \\ (1-\pi) \underbrace{f_N\left(\mu \mid 0, (h^*)^{-1}\right)}_{\mu \sim N(0, \frac{1}{h^*})} & \text{if} \quad \mu \neq 0 \end{array} \right.$$

- Point mass:

$$\frac{\Pr\left(H_1 \mid \bar{Y}\right)}{\Pr\left(H_0 \mid \bar{Y}\right)} = \frac{\int_{\mu \neq 0} f_N\left(\bar{Y} \mid \mu, \sigma^2/T\right) P\left(\mu\right) d\mu}{f_N\left(\bar{Y} \mid \mu = 0, \sigma^2/T\right)}$$

$$= \frac{(1-\pi)}{\pi} \frac{\int \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp\left[-\left(\bar{Y}-\mu\right)^2 \frac{T}{2\sigma^2}\right] \exp\left[-\frac{(\mu)^2 h^*}{2}\right] d\mu}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp\left[-\left(\bar{Y}\right)^2 \frac{T}{2\sigma^2}\right]}$$

- Complete the square in the numerator and integrate out μ
- Side manipulations: Look at numerator

$$\exp\left[-\frac{T}{2\sigma^2}(\bar{Y}^2 - 2\mu\bar{Y} + \mu^2) - \frac{\mu^2}{2}h^*\right]$$

- Complete the square to reach:

$$\exp\left[-\frac{T\bar{Y}^2}{2\sigma^2}\right] \qquad \exp-\left[\left(\frac{h^*}{2}+\frac{T}{2\sigma^2}\right)\mu^2-\frac{2T\bar{Y}}{2\sigma^2}\mu\right]$$

$$=\left(h^*+\frac{T}{\sigma^2}\right)^{-\frac{1}{2}}\sqrt{2}\pi\exp\left[-\frac{1}{2}\frac{\left(\frac{T\bar{Y}}{\sigma^2}\right)^2}{\left(\frac{T}{\sigma^2}+h^*\right)}\right]\cdot$$

$$\exp\left[-\frac{T\bar{Y}^2}{2\sigma^2}\right]\frac{\left(h^*+\frac{T}{\sigma^2}\right)^{\frac{1}{2}}}{\sqrt{2}\pi}\cdot$$

$$\cdot\exp\left[-\frac{1}{2}\left(h^*+\frac{T}{\sigma^2}\right)\left[\mu^2-\left(\frac{\frac{2T\bar{Y}}{\sigma^2}}{\frac{T}{\sigma^2}+h^*}\right)\mu+\left(\frac{\frac{T\bar{Y}}{\sigma^2}}{\frac{T}{\sigma^2}+h^*}\right)^2\right]\right]$$

- Then integrate out the μ (using a conjugate prior) and we get (cancelling terms):

$$\frac{P\left(H_1 \mid \bar{Y}\right)}{P\left(H_0 \mid \bar{Y}\right)} = \left[\frac{1-\pi}{\pi}\right]\left(h^* + \frac{T}{\sigma^2}\right)^{-\frac{1}{2}}$$

$$\cdot \exp\left[\left(\frac{T\bar{Y}^2}{\sigma^2}\right)\left(\frac{1}{2}\right)\left(\frac{\frac{T}{\sigma^2}}{\frac{T}{\sigma^2}+h^*}\right)\right]$$

$$= \left[\frac{1-\pi}{\pi}\right]\underbrace{\left(1+\frac{T}{h^*\sigma^2}\right)^{-\frac{1}{2}}\exp\left[\left(\frac{\bar{Y}}{\frac{\sigma}{\sqrt{T}}}\right)^2\left(\frac{1}{2}\right)\left(\frac{1}{1+\frac{\sigma^2 h^*}{T}}\right)\right]}_{\text{Bayes factor}}$$

$$= \frac{1-\pi}{\pi} \left(\frac{1}{1+\frac{T}{h^{*}\sigma^{2}}} \right)^{\frac{1}{2}} \exp \left[\frac{t^{2}}{2} \left(\frac{1}{1+\frac{\sigma^{2} h^{*}}{T}} \right) \right]$$

- Notice that the higher $\left(\frac{\bar{Y}}{\frac{\sigma}{\sqrt{T}}} \right) =$ "t", the more likely we reject H_0.
- However, as $T \to \infty$, for fixed "t", we get $\dfrac{\Pr\left(H_1 \mid \bar{Y}\right)}{\Pr\left(H_0 \mid \bar{Y}\right)} \to 0$.
- Notice "t"$=\sqrt{T}\frac{\bar{Y}-\mu}{\sigma}$ for $\mu = 0$; this is $O_P(1)$.
- $\therefore$ we support H_0 ("Lindley Paradox")

- Bayesians use sample size to adjust “critical region” or rejection region.
- In classical case, we have that with α fixed, the power of the test goes to 1. (It overweights the null hypothesis.)
- Issue: which weighting of α and β is better?