Gov 2001 Section 8, 2025

Hypothesis Testing

1. For a parameter θ, the null and alternative hypotheses refer to a partition of parameter space Θ into two disjoint parts Θ_0 and Θ_1, where $\Theta = \Theta_0 \cup \Theta_1$. We write

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

   Note: We want to prove the alternative and reject the null.

2. One-sided vs. two-sided:

   (a) One-sided: $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ or $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$. In this case, Θ is partitioned at a single point θ_0.

   (b) Two-sided: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. In this case, the null parameter space is a single point, i.e., $\Theta_0 = \{\theta_0\}$.

3. We cannot observe true θ, so like in point estimation, we use the data or statistic to infer θ. The *rejection region* R of a hypothesis test is a set of possible values of the data y, where we reject H_0 if $y \in R$ and retain H_0 if $y \notin R$. Equivalently, to simplify the test, we can define the rejection region in terms of a test statistic $T(Y)$, such that $R = \{y : T(y) > c\}$ or $R = \{y : T(y) < c_L \text{ or } T(y) > c_U\}$, where c, c_U, c_L are called critical values. A hypothesis testing procedure (or test) specifies which values of the data (or which values of the *test statistic*, as a function of the data) lead to H_0 being rejected or retained.

4. There are two types of errors in hypothesis testing:

   - Type I error (False positive): $\theta \in \Theta_0$, but $y \in R$
   - Type II error (False negative): $\theta \in \Theta_1$, but $y \notin R$

   We often care more about Type I error and try to construct hypothesis tests in a way that controls the probability of Type I error occurring.

5. The *power* function of a test is defined as:

$$\beta(\theta) = P(Y \in R \mid \theta)$$

   In words, this measures how likely we are to reject the null under a given value of θ. Usually, we calculate $P(Y \in R \mid \theta \in \Theta_1)$, i.e., the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true. We want to maximize the power. A *power analysis* evaluates this function for different sample sizes to find the optimal n.

6. The *size* (or level) of a test is the maximum probability of Type I error occurring:

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta) = \max_{\theta \in \Theta_0} P(Y \in R \mid \theta)$$

To construct a hypothesis test that controls Type I error, we set some value of α (prior to seeing the data, most frequently $\alpha = 0.05$, but achieving a smaller α is always better), and then define the rejection region R such that the (maximum) probability of Type I error occurring is equal to α. We call this an α-sized test. We want to minimize the size.

7. Given data y, the *p-value* is the smallest α at which we can reject H_0:

$$p(y) = \min\{\alpha : T(y) \in R_\alpha\}$$

where R_α is the rejection region for a test with size α. For a certain level α, if $p(y) < \alpha$, we reject H_0 at level α; otherwise, we do not reject. This is why we often use p-value as an indicator of statistical significance in social science research.

Another interpretation of the p-value is the probability of observing more extreme data (or the test statistic value) than the current y or $T(y)$ if H_0 is true. For example, let the rejection region be $R_\alpha = \{y : T(y) \geq c_\alpha\}$, and by the definition of size α, we have

$$p(y) = \min_{T(y)\in R_\alpha} \alpha = \min_{T(y)\geq c_\alpha} P(T(Y) \geq c_\alpha \mid \theta \in \Theta_0) = P(T(Y) \geq T(y) \mid \theta \in \Theta_0)$$

Note: The p-value is *not* the probability that H_0 is true. Large p-values could mean either (1) H_0 is true, or (2) the test has low power.

8. To construct a hypothesis test, we need to specify both the null hypothesis (and alternative) and the rejection region:

   (a) Based on the scientific question, define the null and alternative hypotheses H_0 and H_1 (e.g. one-sided or two-sided) before observing the data.

   (b) Choose a test statistic $T(Y)$ and find its distribution under the null, i.e. $T(Y)|(\theta = \theta_0)$.

   (c) Determine the rejection region $R = \{y : T(y) > c\}$ (I use a one-sided test as an example; R can be in other forms). This is usually done by choosing c to obtain an α-sized test, that is, controlling the Type I error rate such that

$$P(T(Y) \in R \mid \theta = \theta_0) = P(T(y) > c \mid \theta = \theta_0) \leq \alpha.$$

   Remarks:

   - Choosing $T(Y)$ can be tricky. This is usually based on some estimator for the parameter θ and finding some pivot.
   - Rejection region for a one-sided test: $R = \{y : T(y) > c\}$ or $R = \{y : T(y) < c\}$; for two-sided:
$$R = \{y : T(y) < c_L \text{ or } T(y) > c_U\}.$$
   - Can also use asymptotic distribution of $T(Y)$ under null if sample size n is large.

Example: We toss a coin n times and observe the data $Y = Y_1, \ldots, Y_n$. We assume a model that the tosses are independent and identical trials where $Y_i \mid p \sim \text{Bern}(p)$. Construct a hypothesis test to determine whether the coin is fair.

Let the true probability of heads be p. Following the steps outlined above:

(a) We conduct a two-sided test: $H_0 : p = \frac{1}{2}$ vs. $H_1 : p \neq \frac{1}{2}$.

(b) A sensible approach is to look at how far is the average of the tosses from 0.5. So we choose the test statistic $T(y) = \bar{Y}$. Since $Y_i \mid p \sim \text{Bern}(p)$, we use the story of the Binomial to get that

$$T(Y) \mid p \sim \frac{1}{n}\text{Bin}(n, p).$$

(c) We want to reject H_0 if the average is too far from 0.5, so we define our rejection region in the form

$$R = \{y : T(y) < c_L \text{ or } T(y) > c_U\}.$$

Now, controlling for the Type I error rate at α, we find c_L and c_U by solving:

$$P(\text{reject } H_0 \mid H_0 \text{ true}) = P\left(y \in R \,\middle|\, p = \frac{1}{2}\right) = P\left(T(y) < c_L \text{ or } T(y) > c_U \,\middle|\, p = \frac{1}{2}\right) \leq \alpha.$$

Concepts in hypothesis testing are very confusing. Don't worry if you cannot remember or use them correctly for now. You can learn from examples or practice Qs.

Common Hypothesis Tests

1. z-test

   - Construct the test statistic based on CLT and normal approximation/asymptotics. Under H_0, the test statistic is:

   $$T(Y) = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}} \dot{\sim} \mathcal{N}(0, 1)$$

   where $\hat{\theta}$ is a consistent estimator for θ and $\hat{\sigma}$ is a consistent estimator of the standard deviation of $\sqrt{n}\hat{\theta}$.

2. t-test

   - Suppose:
     (a) $\hat{\theta} \sim \mathcal{N}(\theta_0, \sigma^2)$ under H_0
     (b) $\hat{\sigma}^2 = s^2(n-1) \sim \sigma^2\chi^2(n-1)$
     (c) $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$ under H_0
   - Then the test statistic:

   $$T(Y) = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n-1}} \sim t_{n-1}$$

   where t_{n-1} is the student's t-distribution with $n-1$ degrees of freedom.

- Remarks:
  - The normality $\hat{\theta} \sim \mathcal{N}(\theta_0, \sigma^2)$ is true for finite sample size n, not asymptotically. Therefore, t-tests (if applicable) are common for small sample size.
  - The independence condition is necessary!
  - Common settings: normal observations testing mean μ ($\hat{\sigma}^2$ is the sample variance), Linear regression model with test on intercept, coefficients, conditional mean, i.e., $E[Y|X = x]$, or new response ($\hat{\sigma}^2$ is the unbiased estimator for variance of error ϵ), etc.

Constructing Confidence Intervals by Inverting Hypothesis Tests

Suppose that for each $\theta \in \Theta$ we have a hypothesis test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with Type I error equal to α. For each of these hypothesis tests, denote the complement of the rejection region $R^c_{\theta_0,\alpha}$ (recall this is a set of $\mathbf{y}$ values). Then the set

$$C_{\mathbf{Y},\alpha} = \{\theta : \mathbf{Y} \in R^c_{\theta,\alpha}\}$$

is a $100(1-\alpha)\%$ confidence region for the unknown θ.

- *Hypothesis tests correspond one-to-one with confidence intervals.*
- *If a CI contains θ_0 then we would not reject in the corresponding hypothesis test.*
- *If we do not reject $H_0 : \theta = \theta_0$ then the corresponding CI contains θ_0.*

Practice Questions

1. A welfare policy is applied to a group of n people to counter job loss caused by trade shocks, resulting in $X_1, \ldots, X_n$ which are i.i.d. $\mathcal{N}(\theta, 1)$, where θ is the theoretical mean effect of the remedy (e.g., the logged wage difference between new and old jobs), defined so that $\theta > 0$ if the remedy is helpful on average, $\theta < 0$ if it is harmful on average, and $\theta = 0$ if it does nothing. Let our Type I error rate be $\alpha = 0.05$. Consider testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let the rejection region be

   $$\mathcal{R} = \{\mathbf{x} : |\hat{\theta}_{PI}| > c\},$$

   where $\hat{\theta}_{PI}$ is the plug-in estimator of θ.

   (a) What is $\hat{\theta}_{PI}$, the plug-in estimator of θ? Your answer should be a function of the data $X_1, X_2, \ldots, X_n$. Also, under the null, what is the distribution of the estimator? (Hint: Under the null means that we assume that $\theta = 0$.)

   (b) Find c so that the test has Type I error rate (i.e., size) α. (Note that c will depend on the sample size, n.) (Hint: Type I error rate α should be equal to $P(|\hat{\theta}_{PI}| > c)$.)

(c) Find the power of the test, $\beta(\theta)$, for $\theta > 0$. (Hint: This question follows almost the same procedure as parts (a) and (b), except for (1) we don't assume that the null is true and (2) the power function is defined as a conditional probability given θ.)

(d) Prove that the power $\beta(\theta) \to 1$ as $n \to \infty$ (Hint: Use the result from part (c).)

(e) Suppose now that $n = 10^4$ and we observe $\bar{x} = 0.02$. What is the p-value? Does the test say to reject or retain H_0? (Hint: The critical value is the one obtained in part (b), i.e., under the null.)

Solutions:

(a) The plug-in estimator $\hat{\theta}_{PI}$ is the sample mean:

$$\hat{\theta}_{PI} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, by the properties of normal distribution, we have:

$$\bar{X} \sim \mathcal{N}(\theta, \tfrac{1}{n}).$$

Under the null hypothesis $H_0 : \theta = 0$, it follows that:

$$\bar{X} \sim \mathcal{N}(0, \tfrac{1}{n}).$$

(b) We want:

$$\mathbb{P}(|\bar{X}| > c \mid H_0) = \alpha.$$

Under H_0, $\bar{X} \sim \mathcal{N}(0, \frac{1}{n})$. So, by symmetry of normal,

$$\mathbb{P}(|\bar{X}| > c) = 2 \cdot \mathbb{P}(\bar{X} > c) = \alpha.$$

Standardizing:

$$\mathbb{P}\left(\frac{\bar{X}}{1/\sqrt{n}} > c\sqrt{n}\right) = \mathbb{P}(Z > c\sqrt{n}) = 1 - \Phi(c\sqrt{n}) = \frac{\alpha}{2}, \quad Z \sim \mathcal{N}(0,1)$$

Thus,

$$c = \frac{z_{1-\alpha/2}}{\sqrt{n}},$$

where $z_{1-\alpha/2}$ is the standard normal critical value. For $\alpha = 0.05$, $z_{0.975} \approx 1.96$, so

$$c = \frac{1.96}{\sqrt{n}}.$$

(c) Under the alternative $X_i \sim \mathcal{N}(\theta, 1)$, so $\bar{X} \sim \mathcal{N}(\theta, \frac{1}{n})$.
The power is the probability of rejecting the null when $\theta \neq 0$, i.e.:

$$\beta(\theta) = \mathbb{P}(|\bar{X}| > c \mid \theta).$$

Standardizing (notice that the distribution of $\bar{X}$ is no longer symmetric to 0):

$$\beta(\theta) = \mathbb{P}\left(\bar{X} < -c \mid \theta\right) + \mathbb{P}\left(\bar{X} > c \mid \theta\right) = \mathbb{P}\left(Z < \frac{-c-\theta}{1/\sqrt{n}}\right) + \mathbb{P}\left(Z > \frac{c-\theta}{1/\sqrt{n}}\right),$$

where $Z \sim \mathcal{N}(0,1)$. Simplified:

$$\beta(\theta) = \Phi\left(\sqrt{n}(-c-\theta)\right) + 1 - \Phi\left(\sqrt{n}(c-\theta)\right).$$

(d) As $n \to \infty$, $c = \frac{1.96}{\sqrt{n}} \to 0$. So for fixed $\theta > 0$, we have

$$\sqrt{n}(-c-\theta) \to -\infty, \quad \sqrt{n}(c-\theta) \to -\infty$$

And the difference is $2\sqrt{n}c = 3.92$, which is negligible in relation to infinity. Since Φ is continuous, we have

$$\beta(\theta) = \Phi\left(\sqrt{n}(-c-\theta)\right) + 1 - \Phi\left(\sqrt{n}(c-\theta)\right) \to 0 + 1 = 1.$$

This proves that power goes to 1 as $n \to \infty$.

(e) From part (b), for $n = 10^4$, we have:

$$c = \frac{1.96}{\sqrt{10^4}} = \frac{1.96}{100} = 0.0196.$$

Given $\bar{x} = 0.02$, we compare it directly to the rejection region $\mathcal{R} = \{\bar{x} : |\bar{x}| > 0.0196\}$. Since:

$$|\bar{x}| = 0.02 > 0.0196 = c,$$

We conclude that the result falls in the rejection region. Therefore, we **reject** H_0.

Optionally, to compute the p-value:

$$T = \frac{0.02}{1/\sqrt{10000}} = \frac{0.02}{0.01} = 2,$$

$$p = 2(1 - \Phi(2)) \approx 2(1 - 0.9772) = 0.0456.$$

Since $p < \alpha = 0.05$, we **reject** the null hypothesis.

2. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ known and σ^2 unknown.

Consider the test:

$$H_0 : \sigma^2 \leq 1 \quad \text{vs.} \quad H_1 : \sigma^2 > 1.$$

Adopt the level $\alpha = 0.05$.

(a) First, let's use $T_1(X) = \sqrt{n}\bar{X}$ as the test statistic. Find the distribution of T_1 under $\sigma^2 = 1$. Construct a test based on rejection region $R_1 = \{x : T_1(x) > c_1\}$. Can you comment on the power of the test intuitively?

(b) One way to improve the test in (a) is to construct a test based on $R_2 = \{x : T_1^2(x) > c_2\}$. Find the critical value c_2.

(c) Find the distribution of another test statistic $T_3 = \sum_{i=1}^{n} X_i^2$ under $\sigma^2 = 1$ and construct a test based on it.

(d) Compare the three tests. Derive their corresponding power functions and compare the power curves based on $n = 10, 100, 1000$.

Solutions:

(a) Under the null, with $\sigma^2 = 1$:

$$T_1 \sim \mathcal{N}(0, 1).$$

The Type I error:

$$P(T_1 \in R_1 \mid \sigma^2 = 1) = 1 - \Phi(c_1) = \alpha,$$

where Φ is the CDF for the standard Normal. So $c_1 = \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.95)$.

The power should not be very high because under H_1, when σ^2 is large, there is high probability to observe very negative T_1 but still not reject H_0 by the test.

(b) Under the null, with $\sigma^2 = 1$:

$$T_1^2 \sim \chi^2(1).$$

The Type I error:

$$P(T_1^2 \in R_2 \mid \sigma^2 = 1) = 1 - F_{\chi^2(1)}(c_2) = \alpha,$$

where $F_{\chi^2(1)}$ is the CDF for $\chi^2(1)$. So $c_2 = F^{-1}_{\chi^2(1)}(0.95)$, the 95% quantile of $\chi^2(1)$.

(c) With $\sigma^2 = 1$:

$$T_3 = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

Let the rejection region be $R_3 = \{x : T_3(x) > c_3\}$. To achieve size α, choose:

$$c_3 = F^{-1}_{\chi^2(n)}(0.95),$$

where $F^{-1}_{\chi^2(n)}$ is the quantile function of $\chi^2(n)$.

(d) Power functions:

- $T_1 \sim \mathcal{N}(0, \sigma^2)$, so

$$\beta_1(\sigma^2) = P(T_1 > c_1 \mid \sigma^2) = P\left(\frac{T_1}{\sigma} > \frac{c_1}{\sigma}\right) = 1 - \Phi\left(\frac{c_1}{\sigma}\right).$$

- $T_1^2/\sigma^2 \sim \chi^2(1)$, so

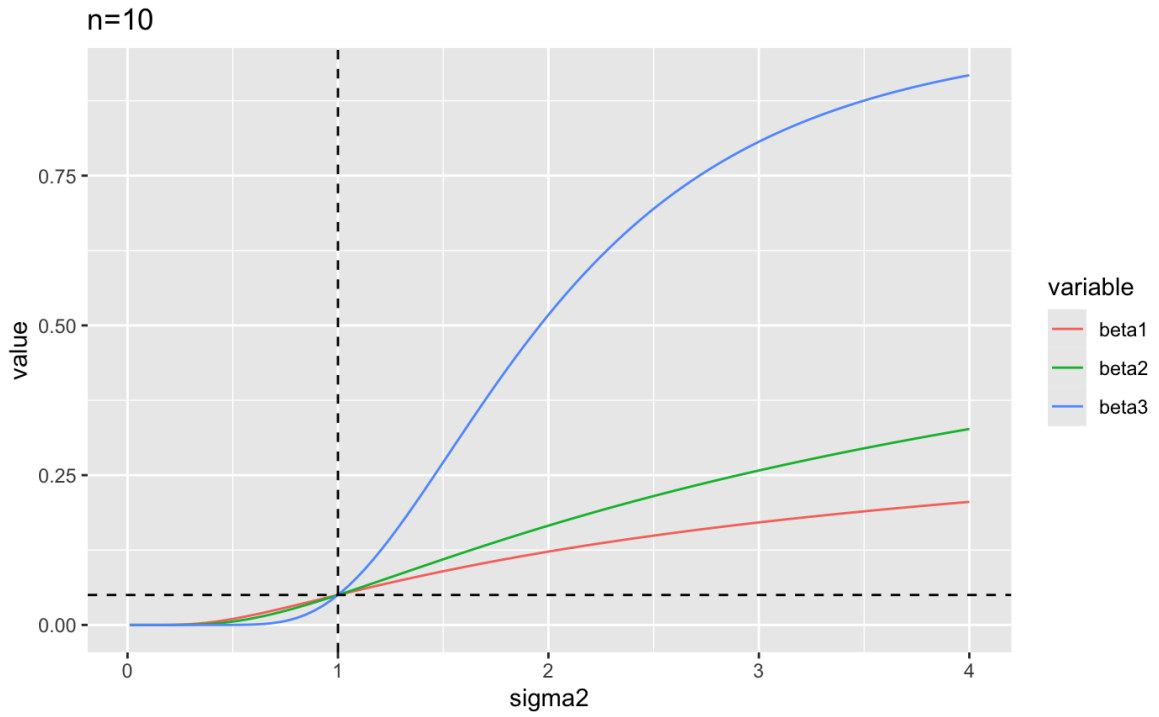
$$\beta_2(\sigma^2) = P(T_1^2 > c_2 \mid \sigma^2) = P\left(\frac{T_1^2}{\sigma^2} > \frac{c_2}{\sigma^2}\right) = 1 - F_{\chi^2(1)}\left(\frac{c_2}{\sigma^2}\right).$$

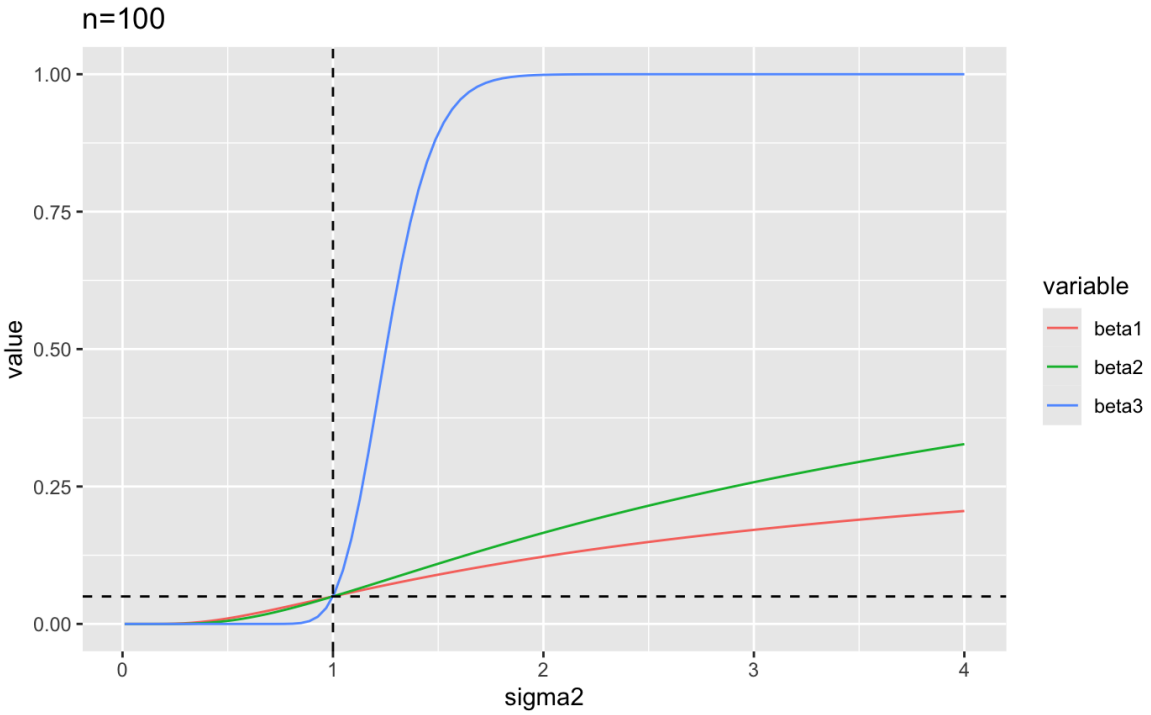
- $T_3/\sigma^2 \sim \chi^2(n)$, so

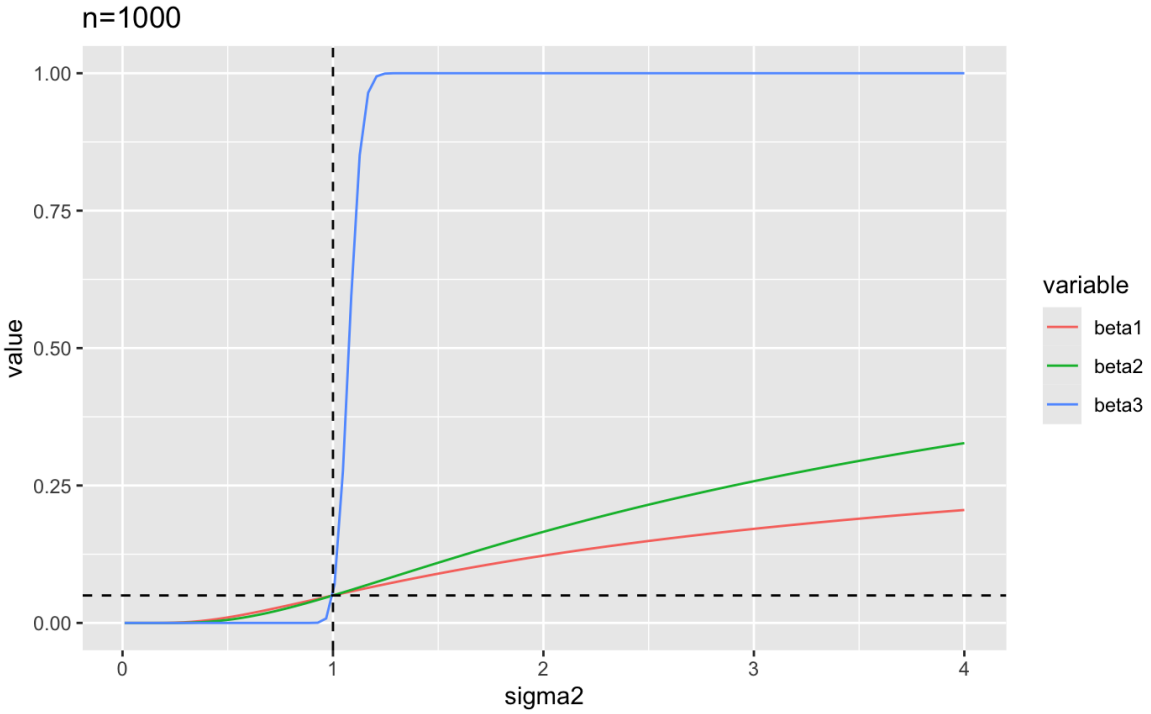
$$\beta_3(\sigma^2) = P(T_3 > c_3 \mid \sigma^2) = P\left(\frac{T_3}{\sigma^2} > \frac{c_3}{\sigma^2}\right) = 1 - F_{\chi^2(n)}\left(\frac{c_3}{\sigma^2}\right).$$

```
library(ggplot2)
library(reshape2)
alpha <- 0.05
n.all <- c(10, 100, 1000)
for(n in n.all){
  sigma2.seq <- seq(0.01, 4, length.out = 101)
  c1 <- qnorm(1 - alpha)
  beta1 <- 1 - pnorm(c1 / sqrt(sigma2.seq))
  c2 <- qchisq(1 - alpha, 1)
  beta2 <- 1 - pchisq(c2 / sigma2.seq, 1)
  c3 <- qchisq(1 - alpha, n)
  beta3 <- 1 - pchisq(c3 / sigma2.seq, n)

  df.power <- data.frame(sigma2 = sigma2.seq, beta1 = beta1, beta2 =
     beta2, beta3 = beta3)
  df.power <- melt(df.power, id.vars = "sigma2")
  g.power <- ggplot(data = df.power, aes(x = sigma2, y = value, color =
      variable)) +
    geom_line() +
    geom_hline(yintercept = 0.05, linetype = 2) +
    geom_vline(xintercept = 1, linetype = 2) +
    ggtitle(paste0("n=", n))
  print(g.power)
}
```




n=100
1.00
0.75
0.50
0.25
0.00
value
0
1
2
3
4
sigma2
variable
beta1
beta2
beta3


n=1000
1.00
0.75
0.50
0.25
0.00
value
0
1
2
3
4
sigma2
variable
beta1
beta2
beta3