第4-3章 假设检验 (p-值)

《统计推断》 第8章

感谢清华大学自动化系江瑞教授提供PPT

有效p-值

- 定义：p-值p(X)是一个满足 $0 \leq p(x) \leq 1, \forall x$ 的检验统计量，如果p(X)的值小可以作为H1的证据。一个p-值成为有效的，如果对于每个 $\theta \in \Theta_0, 0 \leq \alpha \leq 1,$ 都有

$$P_\theta(p(X) \leq \alpha) \leq \alpha$$

- 有效的p-值可以方便地构造出一个水平为 α 的拒绝域

$$R = \{x : p(X) \leq \alpha\}$$

p-value Construction

Let $W(\mathbf{X})$ be a test statistic such that **large** values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

Let $W(\mathbf{X})$ be a test statistic such that **small** values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \leq W(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

Why?

For a certain fixed $\theta \in \Theta_0$. Let $F_\theta(w)$ be the cdf of $W(\mathbf{X})$ and assume that it is a strictly increasing function. Then,

$$p_\theta(\mathbf{x}) = P_\theta(W(\mathbf{X}) \le W(\mathbf{x})) = F_\theta(W(\mathbf{x})).$$

Therefore, $p_\theta(\mathbf{X})$ is equal to the cdf of another random variable $W(\mathbf{X})$!

$$\begin{aligned} P_\theta(p_\theta(\mathbf{X}) \le \alpha) &= P(F_\theta(W(\mathbf{X})) \le \alpha) \\ &= P(F_\theta^{-1}(F_\theta(W(\mathbf{X}))) \le F_\theta^{-1}(\alpha)) \\ &= P(W(\mathbf{X}) \le F_\theta^{-1}(\alpha)) \\ &= F_\theta(F_\theta^{-1}(\alpha)) \\ &= \alpha \end{aligned}$$

Because

$$p(\mathbf{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\mathbf{x}) \ge p_\theta(\mathbf{x}), \quad \text{for every } \mathbf{x},$$

we have

$$P_\theta(p(\mathbf{X}) \le \alpha) \le P_\theta(p_\theta(\mathbf{X}) \le \alpha) = \alpha.$$

Since this is true for every θ, $p(\mathbf{X})$ is a valid p-value.

p-value — the Probability of Observing more Extreme Values Under the Null

Consider the testing problem for normal mean with known variance:

$$H_0 : \mu = \mu_0 \text{ versus } H_0 : \mu > \mu_0$$

The test statistic is

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

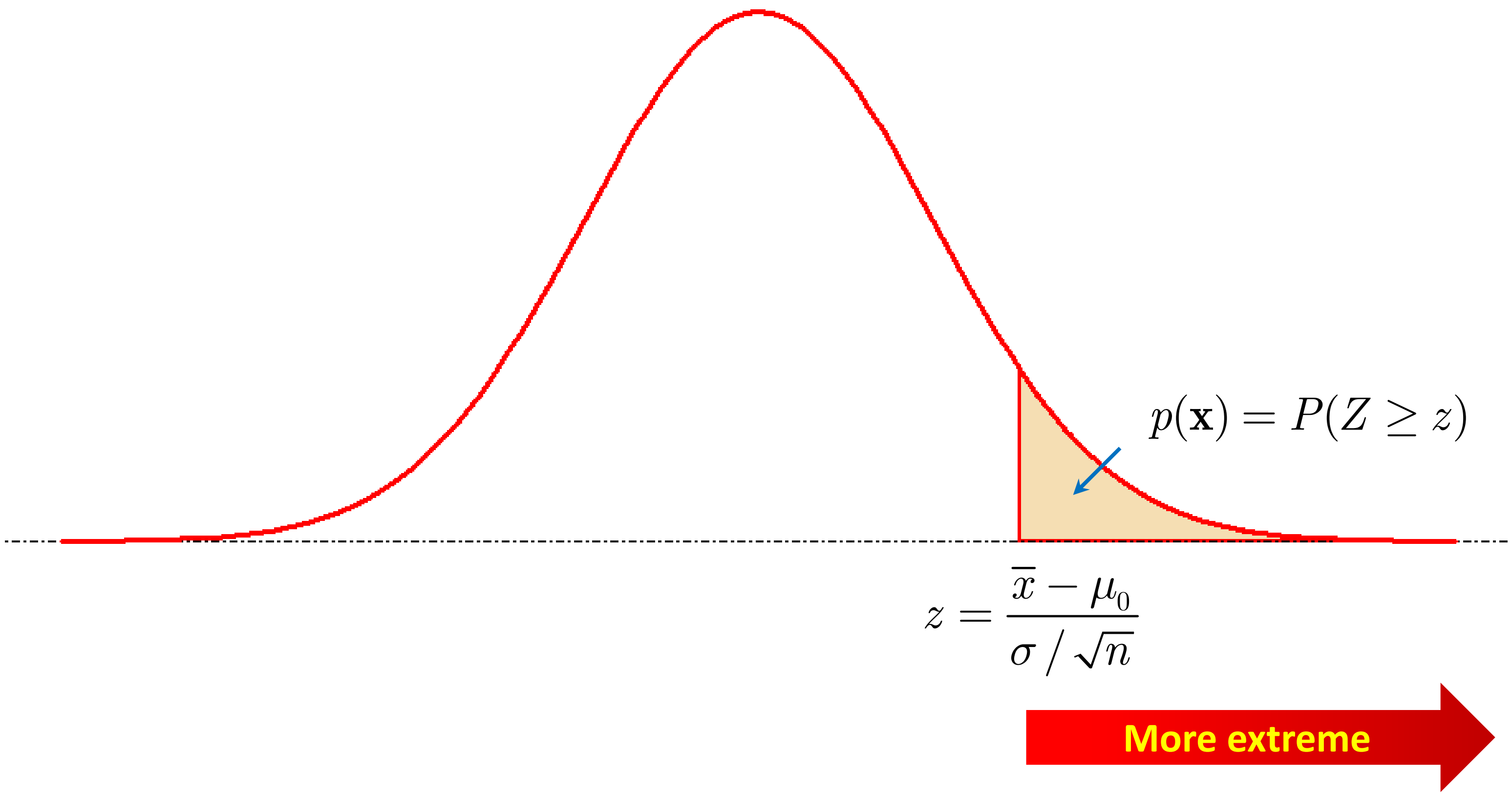
Large values of this statistic give evidence that H_1 is true. When the null is true, this test statistic has a standard normal distribution because $\mu = \mu_0$. Therefore, for a certain observation $\mathbf{x}$,

$$p(\mathbf{x}) = P\left(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \geq \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}\right) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \geq \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} + \frac{\mu_0 - \mu}{\sigma / \sqrt{n}}\right) = P(Z \geq z)$$

which is **the probability of observing more extreme values of the test statistic under the null hypothesis.**

An Illustration

Standard normal



$p(\mathbf{x}) = P(Z \geq z)$

$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

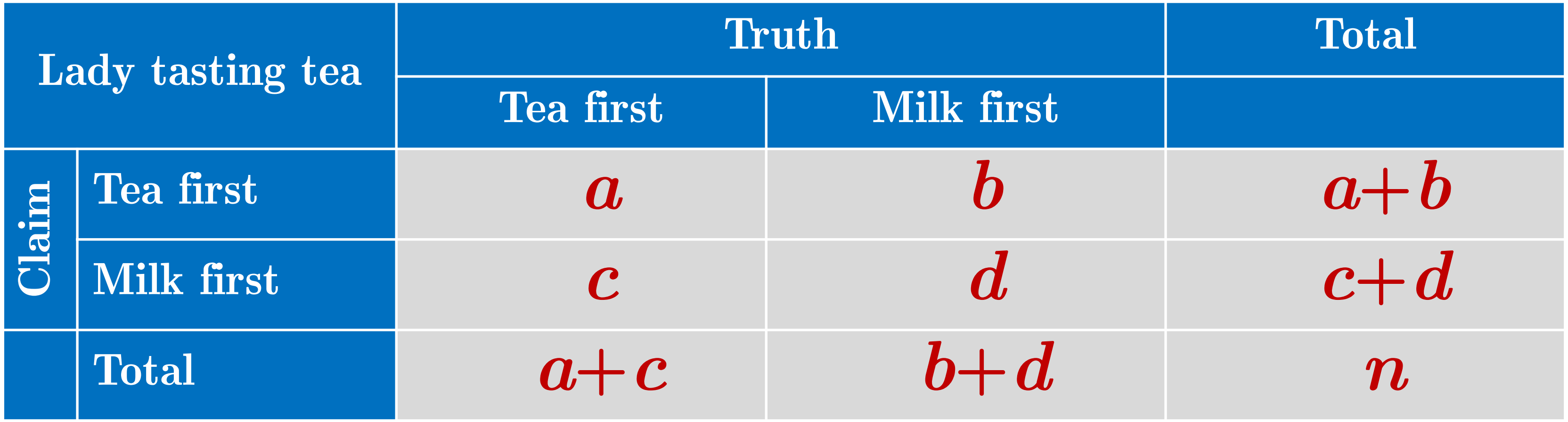
Ronald Aylmer Fisher

Sir Ronald Aylmer Fisher, FRS (February 17, 1890 – July 29, 1962) was an English statistician, evolutionary biologist, eugenicist, and geneticist. Fisher is well known for his contributions to statistics by creating ANOVA (analysis of variance), Fisher's exact test and Fisher's equation.

Anders Hald called him "**a genius who almost single-handedly created the foundations for modern statistical science,**" while Richard Dawkins named him "**the greatest biologist since Darwin.**"

Lady Tasting Tea

- Ms. B. Muriel Bristol claimed that she could distinguish whether tea or milk was first added into a cup of tea
- Mr. R. A. Fisher thought it was nonsense and liked to test it
- How? **The “lady tasting tea” experiment**
- Give the lady eight cups, four of each variety, in random order
- For each cup, record the order of actual pouring and what the lady says the order is



Lady tasting tea		Truth		Total
		Tea first	Milk first	
Claim	Tea first	a	b	$a+b$
	Milk first	c	d	$c+d$
	Total	$a+c$	$b+d$	n

Hypothesis Testing Problem

- According to the experiment design
  - $a+b=c+d=a+c=b+d = 4$
  - Only one of them is free
- If Ms. B. Muriel Bristol claimed correctly
  - a is expected to be large
  - b is expected to be small
  - $a/(a+c) > b\ /(b+d)$
- If Ms. B. Muriel Bristol claimed wrongly
  - No difference between a and b
  - $a/(a+c) = b\ /(b+d)$

$$\boldsymbol{H}_0 : \boldsymbol{p}_T = \boldsymbol{p}_M \quad \text{versus} \quad \boldsymbol{H}_1 : \boldsymbol{p}_T > \boldsymbol{p}_M$$

Hypothesis Test (Fisher's exact test)

- When the null is true (random case)
  - a follows a hypergeometric distribution

$$P(a) = \binom{4}{a}\binom{4}{4-a}\Big/\binom{8}{4}$$

- When the alternative is true
  - a should be large
  - The probability of observing more extreme value in random case is

$$p = P(a \geq a') = \sum_{a=a'}^{4} P(a) = \sum_{a=a'}^{4} \binom{4}{a}\binom{4}{4-a}\Big/\binom{8}{4}$$

  - The smaller the *p*, the stronger the evidence to reject the null
    - a' = 4, $\mathrm{dhyper}(4{,}4{,}4{,}4) = 0.014286$
    - a' = 3, $\mathrm{dhyper}(3{,}4{,}4{,}4) + \ldots = 0.242857$
    - a' = 2, $\mathrm{dhyper}(2{,}4{,}4{,}4) + \ldots = 0.757143$
    - $\ldots$

Are Two Coins Equally Fair?

In a sample from Tsinghua, we observe 146 boys and 55 girls
In a sample from Beida, we observe 128 boys and 96 girls
Is it reasonable to say that on average Beida has more girls than Tsinghua?

In 1000 tosses of coin 1, 520 heads and 480 tails appear.
In 1200 tosses of coin 2, 680 heads and 520 tails appear.
Is it reasonable to assume that the two coins are equally fair?

Tests of Two Proportions

Protocol

Let Populatoin X and Population Y in a contingency have Bernoulli (p_X) and Bernoulli (p_Y) distributions, respectively. We like to test

$$(1)\quad H_0 : p_X = p_Y \quad \text{versus} \quad H_1 : p_X > p_Y;$$

$$(2)\quad H_0 : p_X = p_Y \quad \text{versus} \quad H_1 : p_X < p_Y;$$

$$(3)\quad H_0 : p_X = p_Y \quad \text{versus} \quad H_1 : p_X \neq p_Y.$$

Tests of Two Proportions

Contingency table

	Population X	Population Y	Total
Successes	S_X	S_Y	$S = S_X + S_Y$
Failures	F_X	F_Y	$F = F_X + F_Y$
Total	n_X	n_Y	$n = n_X + n_Y$

Conditional on a Sufficient Statistic

Conditional on a sufficient statistic to define a p-value

Let $W(\mathbf{X})$ be a test statistic such that **large** values of W give evidence that H_1 is true. Let $S(\mathbf{X})$ be a sufficient statistic for the parameter θ **under the null model**. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = P(W(\mathbf{X}) \geq W(\mathbf{x}) \mid S = S(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

Fisher’s Exact Test

Since $S_X = \sum_{i=1}^{n_X} X_i$ is a sufficient statistic for p_X and $S_Y = \sum_{i=1}^{n_X} Y_i$ is a sufficient statistic for p_Y, we can make the decision with the use of only S_X and S_Y.

Obviously, S_X has a binomial $(n_X,\ p_X)$ distribution and S_Y has a binomial $(n_Y,\ p_Y)$ distribution. When **the null is true** $(p_X = p_Y = p)$, the joint pmf of (S_X, S_Y) is

$$f(s_X, s_Y \mid n_X, n_Y, p) = \binom{n_X}{s_X}\binom{n_Y}{s_Y} p^{s_X + s_Y}(1-p)^{(n_X + n_Y)-(s_X + s_Y)},$$

which suggests that $S = S_X + S_Y$ is a sufficient statistic for p under the null hypothesis.

Fisher’s Exact Test

$$H_0 : p_X = p_Y \quad \text{versus} \quad H_1 : p_X > p_Y$$

When **the alternative is true** and $s = s_X + s_Y$ is fixed, we expect to see a large s_X and a small s_Y. So the count S_X can be used as a test statistic for p_X.

When **the null is true**, we have the familiar thing: n_X red balls and n_Y white balls are mixed in an urn, randomly pick up s of them, what is the probability of observing exactly s_X red balls?

$$f(s_X \mid n_X, n_Y, s) = \binom{n_X}{s_X}\binom{n_Y}{s - s_X}\Big/\binom{n_X + n_Y}{s},$$

that is, S_X has a hypergeometric distribution.

Fisher’s Exact Test

Further, what is the probability of observing S_X values that are at least as extreme as s_X under the null?

$$P(S_X \geq s_X \mid S) = \sum_{k=s_X}^{\min\{n_X, s\}} \binom{n_X}{k}\binom{n_Y}{s-k} \Big/ \binom{n_X + n_Y}{s}$$

Therefore,

$$p(s_X) = P(S \geq s_X \mid s_X + s_Y) = \sum_{k=s_X}^{\min\{n_X, s_X + s_Y\}} \binom{n_X}{k}\binom{n_Y}{s_X + s_Y - k} \Big/ \binom{n_X + n_Y}{s_X + s_Y}$$

This defines the **Fisher's exact test**.

```
> fisher.test(...)
```

Simulation of a *p*-value

Simulation of a p-value

1. Calculate the value of the test statistic using the observation $\mathbf{x}$, obtaining w.
2. Simulate a number of n observations $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_n$ **under the null model** and calculate the value of the test statistic for each of them, obtaining w_1, w_2, …, w_n, respectively.
3. Count the number of occurrence that w_i is at least as extreme as w for $i = 1,2,\ldots,n$. Call this number m. Be careful when translating the word "extreme" to the mathematical symbol "$\geq$" or "$\leq$".
4. Divide m by n, obtaining a simulated p-value.

An Example

Test of the mixture coefficient

$\boldsymbol{H_0}$: $\boldsymbol{\lambda = 0}$ **versus** $\boldsymbol{H_1}$: $\boldsymbol{\lambda > 0}$

1. Estimate λ with the use of the EM algorithm. Calculate the likelihood ratio of 1 Gaussian over 2 Gaussians, obtaining w.
2. Simulate a number of n observations $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_n$ according to the 1 Gsussian model (**the null model**). For each of them, apply the EM algorithm to estimate the λ and calculate the likelihood ratio of 1 Gaussian over 2 Gaussians, obtaining w_1, w_2, …, w_n.
3. Count the number of occurrence that w_i is **less than or equal to** w for $i = 1,2,\ldots,n$. Call this number m.
4. Divide m by n, obtaining a simulated p-value.

p-value — a Uniformly Distributed Random Variable

Consider the testing problem for nomal mean:

$$H_0 : \mu = \mu_0 \text{ versus } H_0 : \mu < \mu_0$$

The test statistic is

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

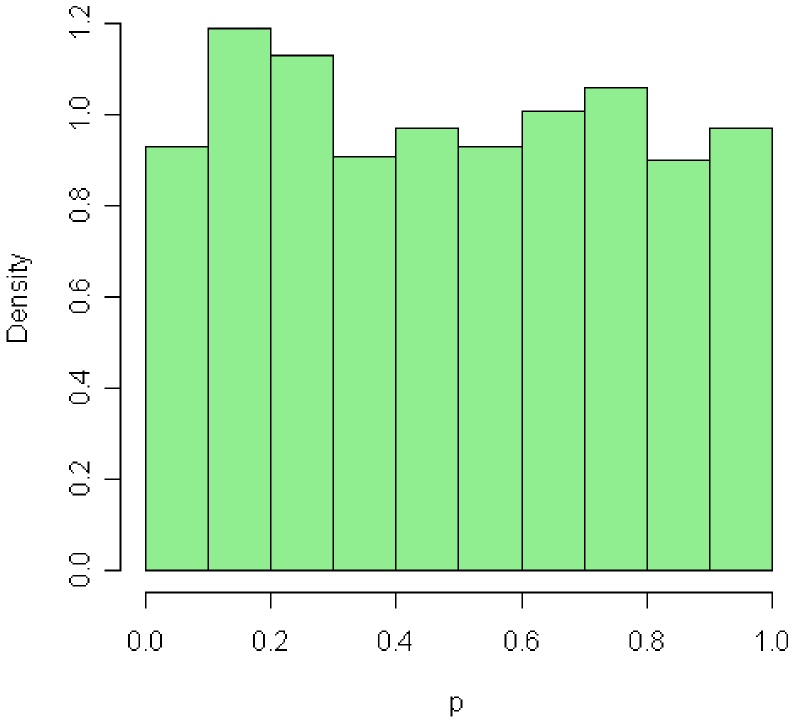
Small values of this statistic give evidence that H_1 is true. When the null is true, this test statistic has a standard normal distribution because $\mu = \mu_0$. Therefore, the p-value is

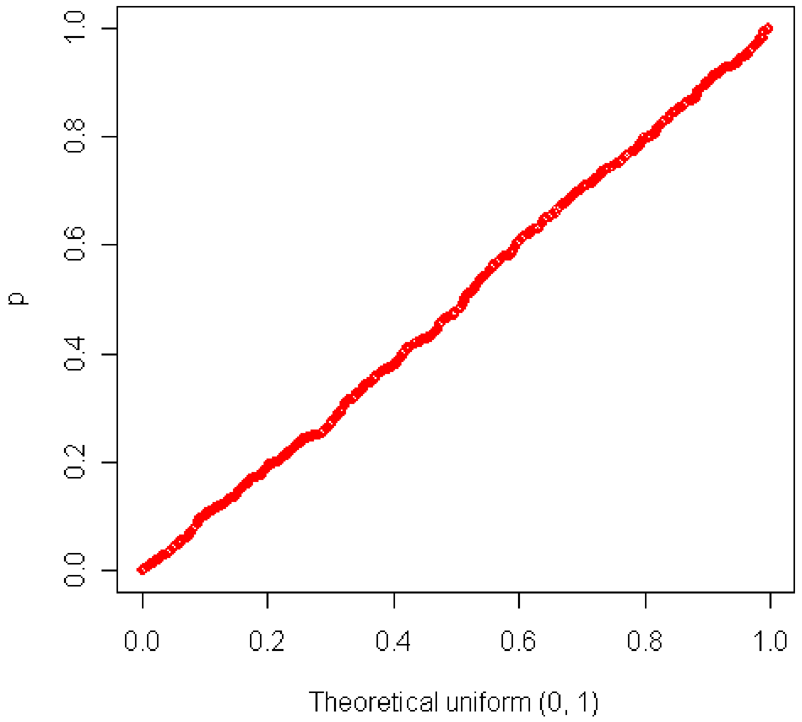
$$p(\mathbf{x}) = P(W(\mathbf{X}) \le W(\mathbf{x})) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} + \frac{\mu_0 - \mu}{\sigma / \sqrt{n}}\right) = P(Z \le z) = \Phi(z),$$

which tells us that $p(\mathbf{X})$ is equal to the standard normal cdf. Therefore, this **p-value has a uniform (0,1) distribution**.

An Illustration

```
n <- 1000; m <- sqrt(n);
p <- rep(0, 1000);
for(i in 1:length(p)){
  x    <- rnorm(n, 0, 1);
  p[i] <- pnorm(m * mean(x));
}
```





Questions

- A *p*-value is a test statistic
- A *p*-value is the probability of observing at least the extreme value as the realized test statistic under the null hypothesis
- A *p*-value is uniformly distributed under the null hypothesis

Hypothesis Testing as a Decision Procedure

- A decision about rejecting or accepting H_0
- Action space
  - a_1 = Reject H_0
  - a_0 = Accept H_0
- What we gain by rejecting or accepting H_0
  - Reduced the parameter space to $\Theta_0{}^c$ (rejecting) or Θ_0 (accepting)
- What we loss by rejecting or accepting
  - The parameter may actually in the opposite space
- The loss reflects the risk of the decision
  - We want to minimize the risk

Loss Function

- The cost of our decision
- 0-1 loss

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}$$

- Generalized 0-1 loss

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_{\mathrm{II}} & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c_{\mathrm{I}} & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}$$

Expected Loss of a Decision

- An action

$$\delta = \delta(\mathbf{x}) = a_0, \text{ or } \; \delta = \delta(\mathbf{x}) = a_1, \mathbf{x} \in \mathcal{X}$$

$$A = \{\mathbf{x} : \delta(\mathbf{x}) = a_0\}, \; R = \{\mathbf{x} : \delta(\mathbf{x}) = a_1\}$$

- δ(**X**) has a Bernoulli distribution

$$P_\theta(\delta(\mathbf{X}) = a_1) = P_\theta(\mathbf{X} \in R) = \beta(\theta)$$

$$P_\theta(\delta(\mathbf{X}) = a_0) = P_\theta(\mathbf{X} \in A) = 1 - \beta(\theta)$$

Expected Loss

- The expected loss of an action

if $\theta \in \Theta_0$,

$$R(\theta, \delta(\mathbf{X})) = \mathrm{E}_\theta L(\theta, \delta(\mathbf{X})) = 0 P_\theta(\delta(\mathbf{X}) = a_0) + P_\theta(\delta(\mathbf{X}) = a_1) = \beta(\theta)$$

if $\theta \in \Theta_0^c$,

$$R(\theta, \delta(\mathbf{X})) = \mathrm{E}_\theta L(\theta, \delta(\mathbf{X})) = P_\theta(\delta(\mathbf{X}) = a_0) + 0 P_\theta(\delta(\mathbf{X}) = a_1) = 1 - \beta(\theta)$$

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_0^c \end{cases} \text{ and } L(\theta, a_1) = \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}$$

Normal Risk Function

Consider the following test of the normal mean with known variance $\sigma^2 = 1$

$$H_0 : \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0$$

LRT suggests a rejection region of

$$R = \left\{ \mathbf{x} : \bar{x} < \mu_0 - z / \sqrt{n} \right\}$$

The power function of such a test is

$$\beta(\mu) = P_\mu \left(\frac{\bar{X} - \mu_0}{1/\sqrt{n}} < -z \right) = P_\mu \left(\frac{\bar{X} - \mu}{1/\sqrt{n}} < -z - \frac{\mu - \mu_0}{1/\sqrt{n}} \right) = \Phi \left(-z - \frac{\mu - \mu_0}{1/\sqrt{n}} \right)$$

The rejection region for test with at most α type I error probability is

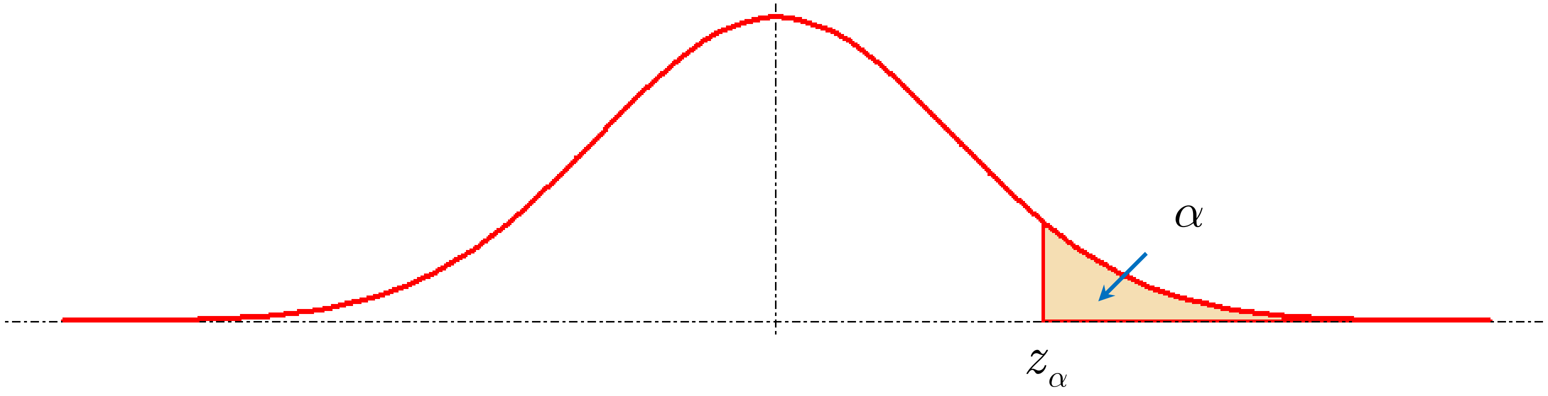
$$R = \left\{ \mathbf{x} : \bar{x} < \mu_0 - z_\alpha / \sqrt{n} \right\}, z_\alpha = \Phi^{-1}(1 - \alpha)$$

The risk function (for 0-1 loss) is

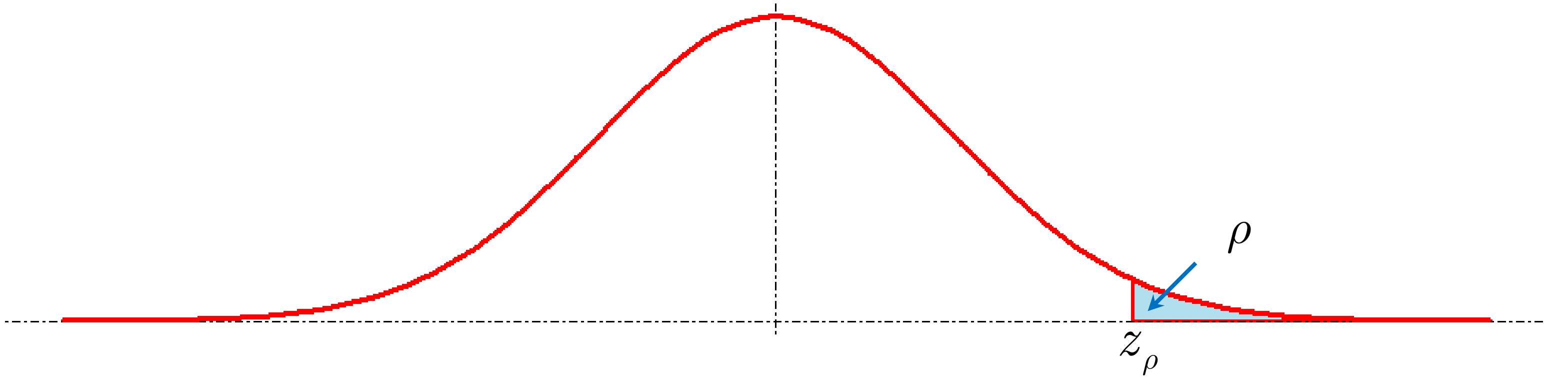
$$R(\mu, \delta) = \begin{cases} \Phi(-z_\alpha - \sqrt{n}(\mu - \mu_0)) & \mu \geq \mu_0 \\ 1 - \Phi(-z_\alpha - \sqrt{n}(\mu - \mu_0)) & \mu < \mu_0 \end{cases}$$

$$P(Z > z_{\alpha}) = \alpha$$

Standard normal

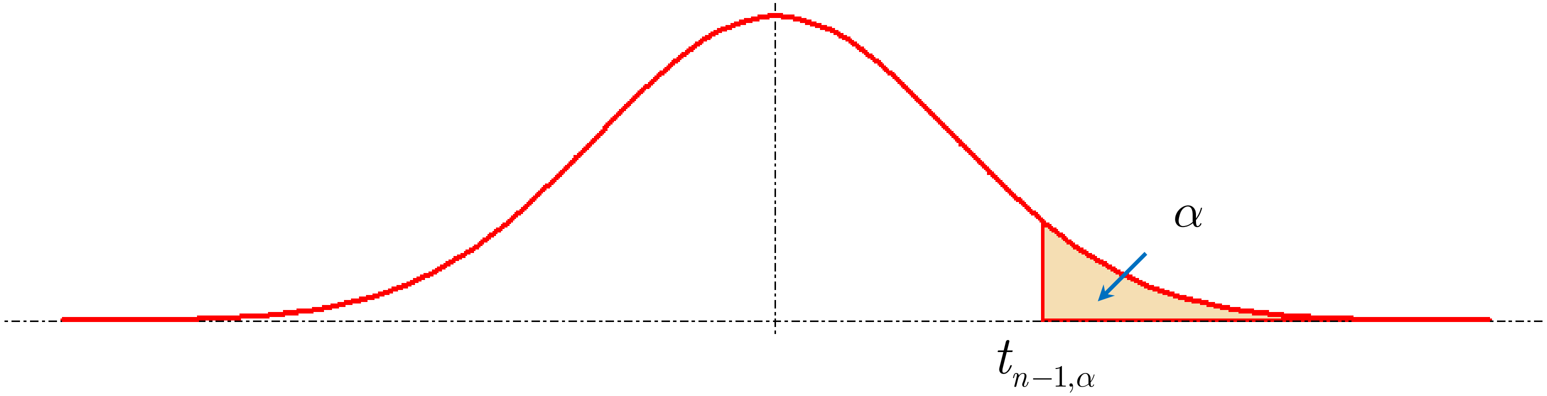


Standard normal

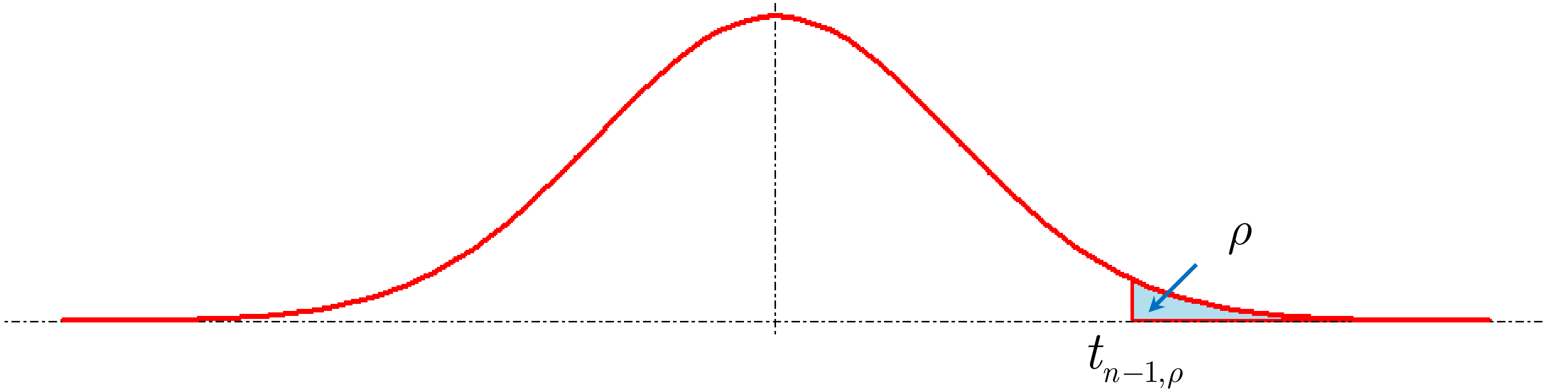


$$P(T_{n-1} > t_{n-1,\alpha}) = \alpha$$

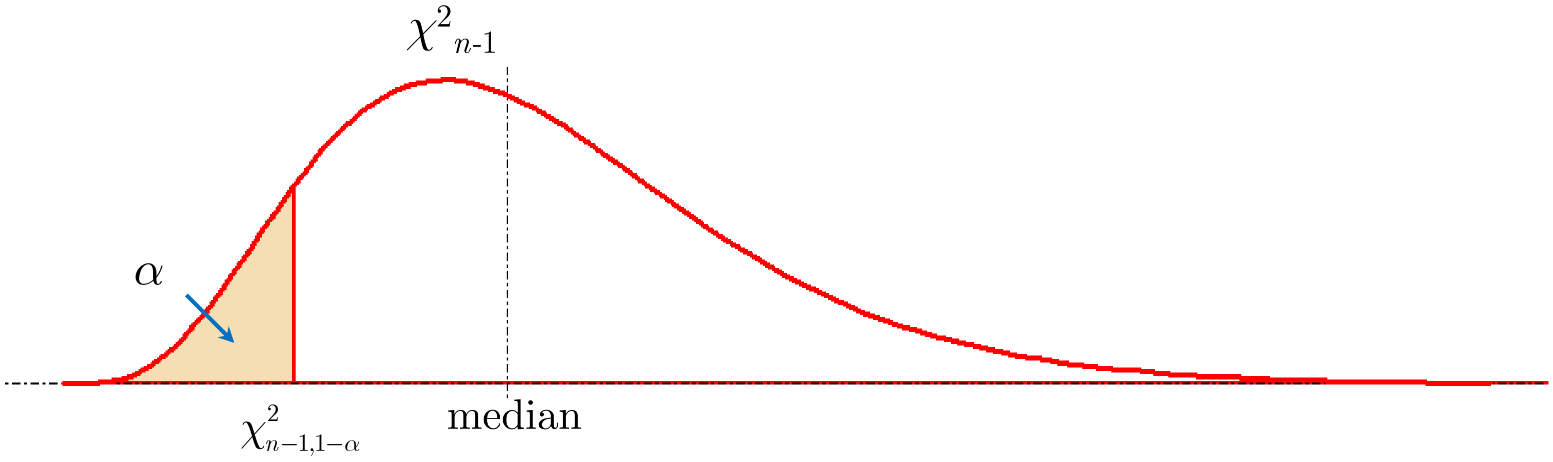
Student's *t* (df = *n*-1)

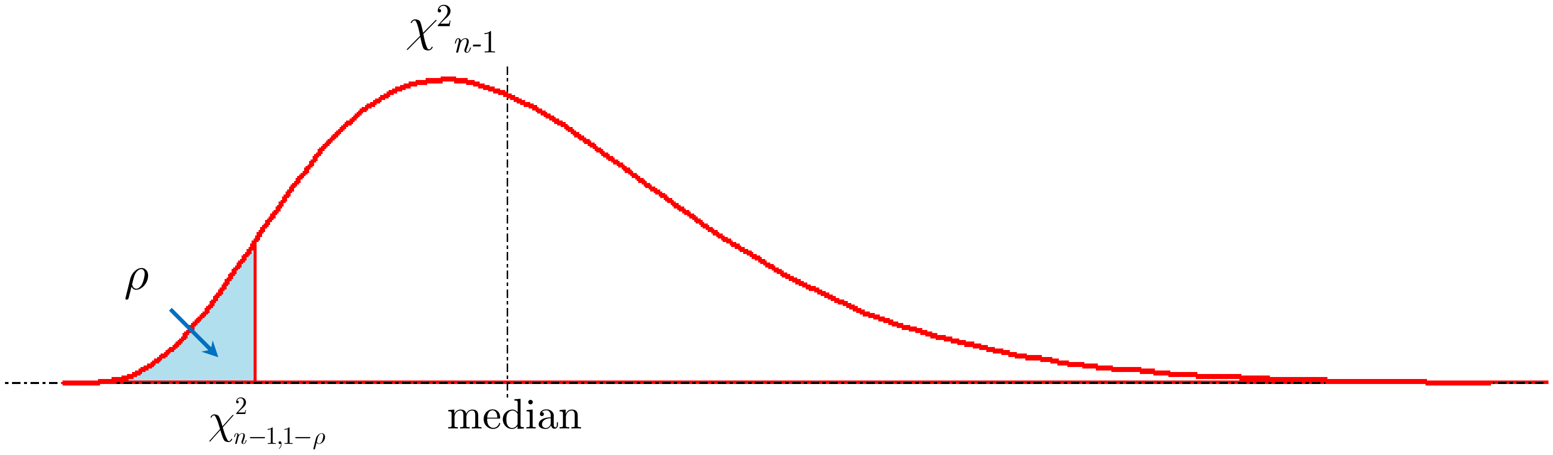


Student's *t* (df = *n*-1)

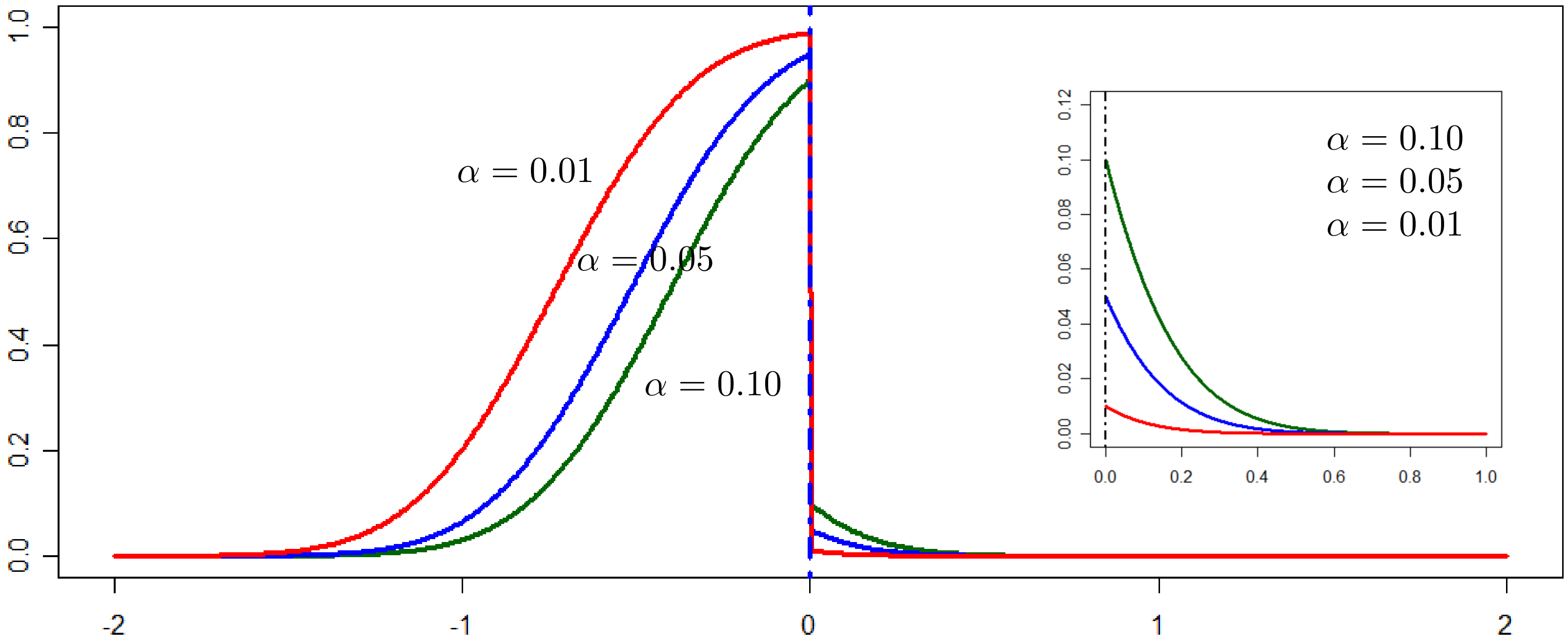


$$P(\chi^2_{n-1} > \chi^2_{n-1,1-\alpha}) = 1 - \alpha$$

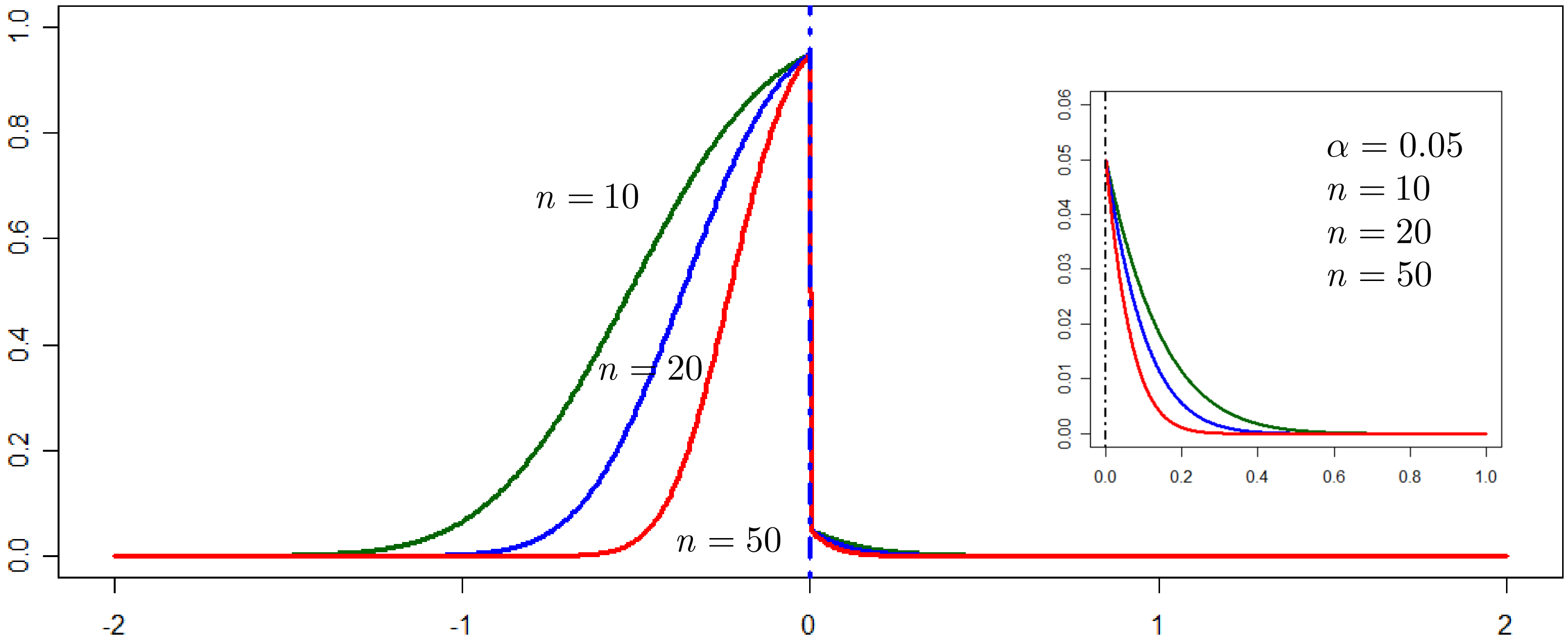




Depends on the Type I Error Probability



Depends on the Size of the Sample


$n = 10$
$n = 20$
$n = 50$
$\alpha = 0.05$
$n = 10$
$n = 20$
$n = 50$