

Statistics 514: Design of Experiments

Topic 2

Topic Overview

This topic will cover

- Basic Statistical Concepts (Montgomery 2-1, 2-2)
- Commonly Used Densities (Montgomery 2-3)

Basic Statistical Concepts

- **Random Variable - Y**

- Quantity (response) capable of taking on a set of values
- Discrete or continuous

$$\sum_i \Pr(Y = y_i) = 1 \text{ or } \int f(y)dy = 1$$

- Described by a probability distribution (density f)

- **Numerical Summaries of a Variable**

- Center - Mean: μ , $E()$
- Spread - Variance: σ^2 , $\text{Var}()$

$$\begin{array}{ll} \mu : & \begin{array}{l} \text{Discrete} \\ \sum y \Pr(Y = y) \end{array} & \begin{array}{l} \text{Continuous} \\ \int y f(y) \end{array} \\ \sigma^2 : & \begin{array}{l} \sum (y - \mu)^2 \Pr(Y = y) \end{array} & \int (y - \mu)^2 f(y) \end{array}$$

- **Independence** – Observations are *statistically independent* if the value of one of the observations does not influence the value of any other observations.

- **Elementary Results of Numerical Summaries**

- $E(aY \pm b) = aE(Y) \pm b$
- $\text{Var}(aY \pm b) = a^2\text{Var}(Y)$
- $E(Y_1 \pm Y_2) = E(Y_1) \pm E(Y_2)$
- $\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2)$.
- If Y_1 and Y_2 are *independent*, $\rightarrow \text{Cov}(Y_1, Y_2) = 0$.
- $\text{Var}(Y) = E(Y^2) - E(Y)^2 = E[(Y - E(Y))^2]$

- $\text{Var}(Y_1 \pm Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2) \pm 2\text{Cov}(Y_1, Y_2)$
- $\text{E}(Y_1 \times Y_2) = \text{E}(Y_1)\text{E}(Y_2)$, if Y_1, Y_2 independent.
- However, $\text{E}\left(\frac{Y_1}{Y_2}\right) \neq \frac{\text{E}(Y_1)}{\text{E}(Y_2)}$.

Common Sample Summaries

- **Sample mean** (\bar{Y})

If Y_1, \dots, Y_n are independent with mean μ and variance σ^2 ,

$$\begin{aligned}\text{E}\left(\frac{1}{n} \sum Y_i\right) &= \frac{1}{n} \sum \text{E}(Y_i) = \frac{1}{n} n\mu = \mu \\ \text{Var}\left(\frac{1}{n} \sum Y_i\right) &= \frac{1}{n^2} \sum \text{Var}(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n\end{aligned}$$

What is the distribution of \bar{Y} ?

$$\begin{aligned}\text{If } Y_i \text{ Normal} &\rightarrow \bar{Y} \text{ Normal} \\ \text{If } Y_i \text{ Other} &\rightarrow \bar{Y} \approx \text{Normal}\end{aligned}$$

The Central Limit Theorem

If Y_1, \dots, Y_n are independent R. V.'s with mean μ and variance σ^2 ,

$$\frac{\sum Y_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0, 1)$$

- **Sample variance** ($S^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$)

$$\begin{aligned}\text{E}(Y_i - \bar{Y}) &= \text{E}(Y_i) - \text{E}(\bar{Y}) = 0 \\ \text{Var}(Y_i - \bar{Y}) &= \text{Var}(Y_i) + \text{Var}(\bar{Y}) - 2\text{Cov}(Y_i, \bar{Y}) \\ &= \sigma^2 + \sigma^2/n - 2\sigma^2/n \\ &= \frac{n-1}{n}\sigma^2 \\ \text{E}(S^2) &= \frac{1}{n-1} \sum \text{Var}(Y_i - \bar{Y}) \\ &= \frac{1}{n-1} n \frac{n-1}{n} \sigma^2 \\ &= \sigma^2\end{aligned}$$

- **Sample standard deviation** ($S = \sqrt{S^2}$)

What is the distribution of S^2 ?

If Y_i Normal, then

$$(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2,$$

where $n-1$ is the degrees of freedom.

Setup

- **Goal:** Learn about population from (randomly) drawn data/sample
- **Model and Parameter:** Assume population (Y) follow a certain model (distribution) that depends on a set of unknown constants (parameters): $Y \sim f(y, \theta)$

Example: Y is the yield of a tomato plant

$$\begin{aligned} Y &\sim N(\mu, \sigma^2) \\ f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ \theta &= (\mu, \sigma^2) \end{aligned}$$

Random sample/observations

- Random sample (conceptual)

$$X_1, X_2, \dots, X_n \sim f(x; \theta)$$

- Random sample (realized/actual numbers)

$$x_1, x_2, \dots, x_n$$

Example: 0.0 4.9 -0.5 -1.2 2.1 2.8 1.2 0.8 0.9 -0.9

Populations/Samples

A *parameter* is the true value of some aspect of the population. (Examples: mean, median, variance, slope)

An *estimator* is a

- statistic that corresponds to a parameter.
- random variable (not based on any particular data).

An *estimate* is a particular value of the estimator, computed from the sample data. It is considered fixed, given the data.

Sampling from a population:

	Collection of possible values	Numerical Summary
What you want to know	Population	Parameter
What you actually get to see	Sample	Statistic

$$\text{Estimator: } \hat{\theta} = g(Y_1, \dots, Y_n)$$

$$\text{Estimate: } \hat{\theta} = g(y_1, \dots, y_n)$$

Example

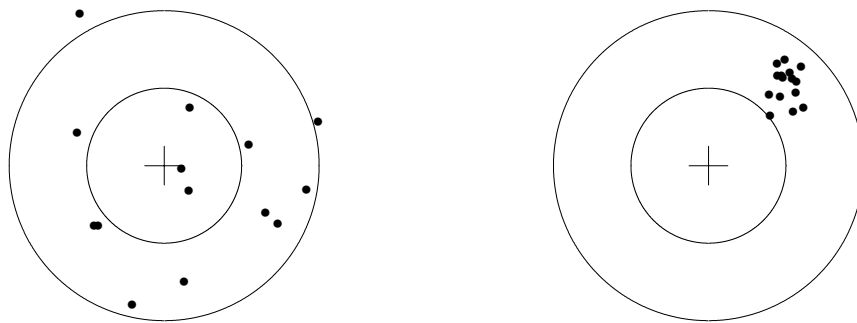
Estimators for μ and σ^2

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \quad ; \quad \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

Estimates

$$\hat{\mu} = \bar{y} = 1.01 \quad ; \quad \hat{\sigma}^2 = s^2 = 3.49$$

Variance vs. Bias



While *variance* refers to random spread, *bias* refers to a systematic drift in an estimator.

Most of the time, we will be concerned with the variance and bias of *estimators*, not populations. (Thus, an estimate of variance may be biased.) Thus, bias and variance are inherent (and thus often subject to manipulation) in a *statistical method*, not the sampling.

An estimator $\hat{\theta}$ of θ is *unbiased* if $E(\hat{\theta}) = \theta$.

Unbiased	Biased
sample mean	sample standard deviation
sample variance	<i>F</i> -ratio

Degrees of Freedom of a sum is equal to the number of elements in that sum that are independent (i.e., free to vary).

For example, if you are told the sum of three elements equals five, you only need to know two of the three elements to know all of them.

General Result:

If Y_i has variance σ^2 and $SS = \sum (Y_i - \bar{Y})^2$ with k degrees of freedom,

$$E(SS/k) = \sigma^2.$$

Sampling/Reference Distribution

Statistical inference/testing: making decision in the presence of variability. Is result of experiment easily explained by chance variation or is it “unusual”?

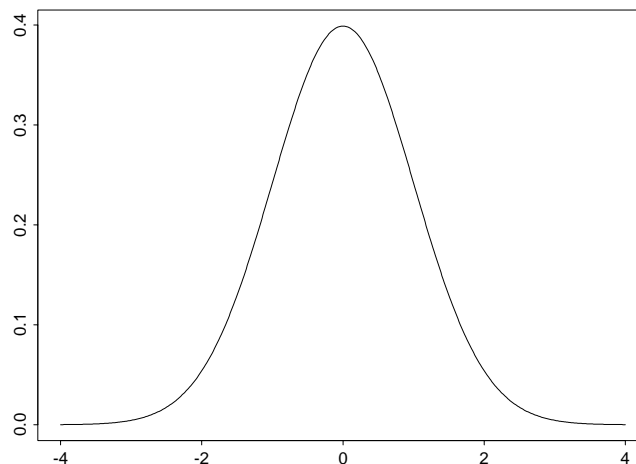
- “Unusual”: Is it unlikely if only chance variation?
- Need distribution of results assuming only chance variation (null distribution)
- Compare observed result with distribution of outcomes
- Example 1: t -test (comparing two means)
 - Calculate observed t -test statistic
 - t distribution summarizes outcomes under Null hypothesis
 - Compare observed result with distribution
- Example 2: randomization test
 - Chance variation due to randomization
 - Generate all possible outcomes (each equally likely)
 - Compare observed result with distribution of outcomes

Standard Normal Distribution

Take $Z_i \sim N(0, 1)$ independent $i = 1, \dots, n$.

So $P(a < Z_i < b) = \int_a^b f(z)dz$, where

$$f(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$



The density of $Z = (Z_1, \dots, Z_n)$ is

$$f(z) = (2\pi)^{-n/2} e^{-\frac{1}{2} \sum_{i=1}^n z_i^2}$$

Normal with mean μ and variance σ^2

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

Standardizing

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Application

Used to *model* random observations.

$$Y = \mu(X) + e\sigma, \text{ where } e \sim N(0, 1). \\ \text{Given } X = x, Y \sim N(\mu(x), \sigma^2).$$

Multivariate Normal

$$X = \mu + AZ \sim N(\mu, A'A). \\ Z = A^{-1}(X - \mu) \sim N(0, I).$$

Special case: A orthogonal, $A'A = I$.

SAS Code

```
data normal;
qtl = probit(0.975); /* to get z value for 95% confidence interval */
pval = 2*(1-probnorm(2.3)); /* to get p-value of 2-sided z-score 2.3 */
run;
```

```
proc print data = normal;
run;
```

Obs	qtl	pval
1	1.95996	0.0214

R Code

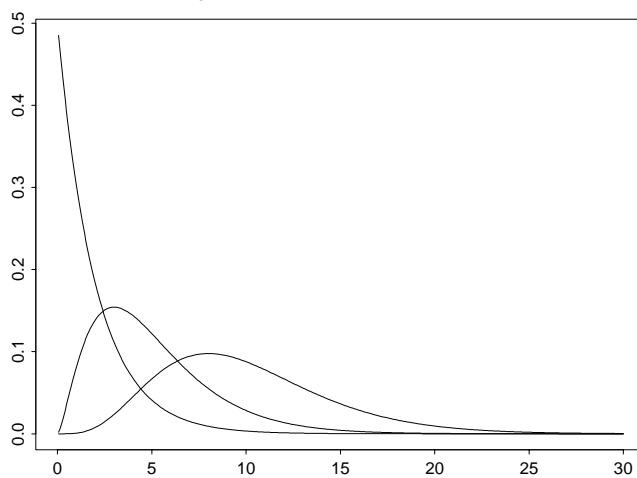
```
> qtl = qnorm(0.975)
> qnorm(0.975)
[1] 1.959964
> 2*(1-pnorm(qtl))
[1] 0.05
```

Chisquare distribution

Chisquare distribution on n degrees of freedom. Sums of squares of normals; usually in variance estimates

$$S^2 = \sum_{i=1}^n Z_i^2 \sim \chi_{(n)}^2$$
$$\sum (Y_i - \bar{Y})^2 / \sigma^2 \sim \chi_{(n-1)}^2$$

Chisquare densities on 2, 5, 10, d. f.



$$E(S^2) = n, \text{Var}(S^2) = 2n$$

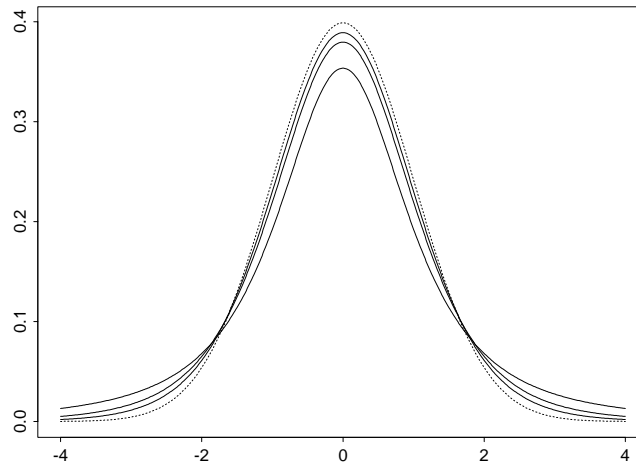
- For large n , $\chi_{(n)}^2 \doteq N(n, 2n)$ by CLT.
- In SAS, use `probchi(q, df)` for p -values and `cinvc(p, df)` for quantiles.
- In R, use `pchisq(q, df)` for p -values and `qchisq(p, df)` for quantiles.

Student's t Distribution

If $X \sim N(0, 1)$ and $s^2 \sim \frac{1}{d}\chi_{(d)}^2$, independent, then $t = \frac{X}{s} \sim t_{(d)}$.

Recipe: $t_{(d)} = N(0, 1) / \sqrt{\text{indep } \chi_{(d)}^2 / d}$.

t densities on 2, 5, 10, ∞ d. f.



small $d \Rightarrow$ heavy tails

Handy theorem

$X_i \sim N(\mu, \sigma^2)$ independent. Let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Then

$$\bar{X} \sim N(\mu, \sigma^2/n), \quad s^2 \sim (n-1)^{-1} \chi_{(n-1)}^2, \quad \text{indep}$$

Sample standardization

So, if $X_i \sim N(\mu, \sigma^2)$ independent, then

$$t = \sqrt{n} \frac{\bar{X} - \mu}{s} \sim t_{(n-1)}$$

- In SAS, use `probt(q , df)` for p -values and `tinvt(p , df)` for quantiles.
- In R, use `pt(q , df)` for p -values and `qt(p , df)` for quantiles.

Fisher's F Distribution

If $SS_N \sim \chi_{(n)}^2$ independent of $SS_D \sim \chi_{(d)}^2$,

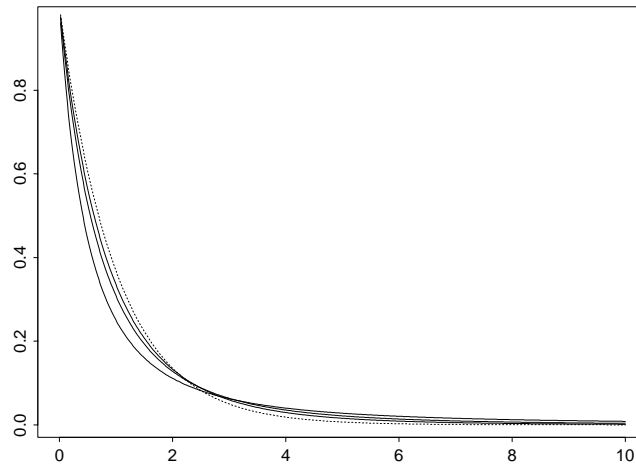
Then

$$F = \frac{\frac{1}{n} SS_N}{\frac{1}{d} SS_D} \equiv \frac{MS_N}{MS_D} \sim F_{n,d}$$

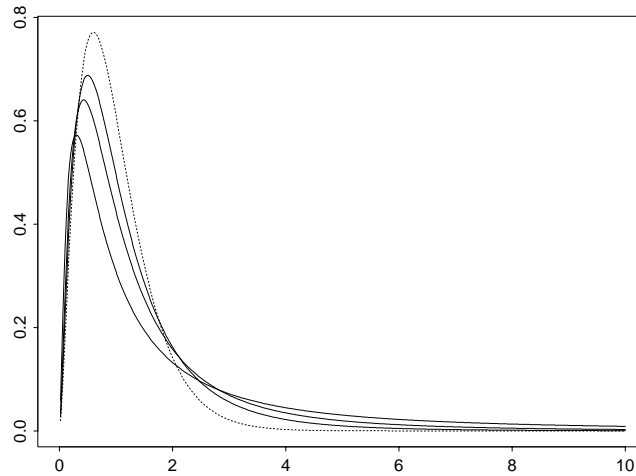
Notes

- $1/F_{n,d} \sim F_{d,n}$
- $F_{1,d} = t_{(d)}^2$
- As $d \rightarrow \infty$, $F_{n,d} \rightarrow \chi_{(n)}^2/n$.

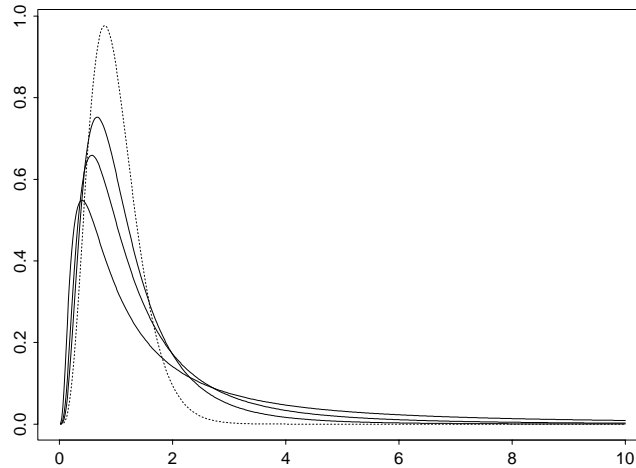
$F_{2,d}$ densities with $d \in \{2, 5, 10, \infty\}$



$F_{5,d}$ densities with $d \in \{2, 5, 10, \infty\}$



$F_{10,d}$ densities with $d \in \{2, 5, 10, \infty\}$



- In SAS, use `probf(q, df1, df2)` for p -values and `finv(p, df1, df2)` for quantiles.
- In R, use `pf(q, df1, df2)` for p -values and `qf(p, df1, df2)` for quantiles.

Noncentral Distributions

Noncentral chisquare

$X_i \sim N(a_i, 1)$ independent

$$C = \sum_{i=1}^n X_i^2 \sim \chi_{(n)}^2(\phi), \phi = \sum_i a_i^2$$

Miracle: only depends on $\|a\|$

Arises: mean square when $\mu \neq 0$ (in alternate hypotheses)

- In SAS, use `probchi(q, df, phi)` for p -values and `cinv(p, df, phi)` for quantiles.
- In R, use `pchisq(q, df, phi)` for p -values and `qchisq(p, df, phi)` for quantiles.

Noncentral F

$$F'_{n,d}(\phi) = \frac{\chi_{(n)}^2(\phi)/n}{\chi_{(d)}^2/d}$$

Arises: probability of significant F -test under alternative (i.e. power of F test).

- In SAS, use `probf(q, df, phi)` for p -values and `finv(p, df, phi)` for quantiles.
- In R, use `pf(q, df, phi)` for p -values.

Noncentral t

$$t'_d(a_1) = \frac{N(a_1, 1)}{\sqrt{\chi^2_{(d)}}}$$

Power of t test

- In SAS, use `probt(q , df , a_1)` for p -values and `tinvt(p , df , a_1)` for quantiles.
- In R, use `pt(q , df , a_1)` for p -values.

Doubly noncentral F

$$F''_{n,d}(\phi) = \frac{\chi^2_{(n)}(\phi_n)/n}{\chi^2_{(d)}(\phi_d)/d}$$

For power when error mean square corrupted.

Noncentral distributions

Widely tabled

Watch parameterization closely

See

1. Encyclopedia of the statistical sciences
2. Johnson and Kotz's books on distributions
3. the Web

Review on Finding p -values

Building block for p -values: *cumulative distribution function* (cdf)

- Let X be a random variable.
- The cdf for X is a function of x such that $cdf(x) = P(X < x)$.
- **Note:** $P(X > x) = 1 - P(X < x)$
- Examples of functions which evaluate cdf for different distributions: `probnorm`, `probchi`, `probt`, `probf`.

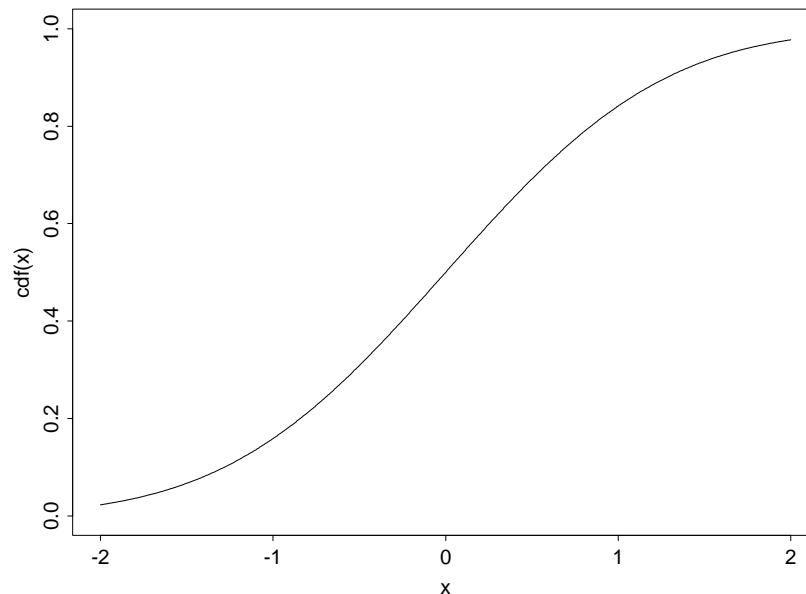


Figure 1: Example of cumulative distribution (for standard normal random variable)

1-sided p -values

- Used in F -tests, t -tests with $>$ or $<$ alternatives (not \neq).
- Procedure: Get test statistic u .
- If alternative hypothesis is $\theta > 0$, then usually p -value is $P(X > u)$.
- *Example*
 - If alternative hypothesis is $\mu_1 - \mu_2 > 0$ and t statistic is 2.5 with 14 degrees of freedom, the p -value is $1 - P(t_{14} < 2.5) = 0.0127$.

2-sided p -values

- Used in t -tests with \neq alternative.
- Procedure: Get test statistic u .
- If alternative hypothesis is $\theta \neq 0$, then usually p -value is usually $P(X > |u|) = 2(1 - \text{cdf}(|u|))$.
- *Example*
 - If alternative hypothesis is $\mu_1 - \mu_2 \neq 0$ and t statistic is 2.5 with 14 degrees of freedom, the p -value is $2(1 - P(t_{14} < 2.5)) = 0.0255$.

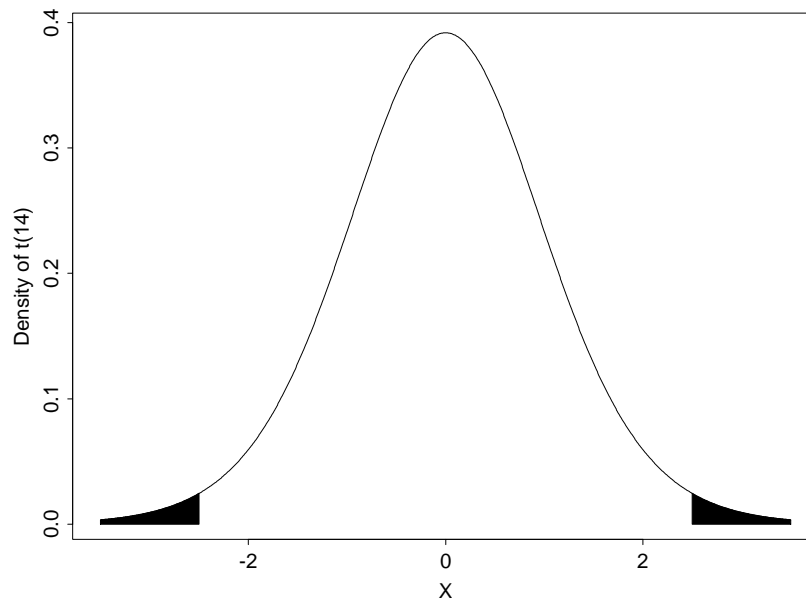


Figure 2: Two-sided p -value corresponds to area in both shaded regions

- *Notes*

- Most of the time, SAS output will report the p -value associated with a particular test. However, this will not always be true.
- Sometimes, a test statistic is so large that the p -value is reported as 0. This, of course, is not true; you should report, in this case, that the p -value is too small to be determined numerically.

Cutoff values/rejection regions

- Often, we report a *cutoff value* for a particular test at a particular α level.
- Test statistics that are larger (usually) than the cutoff value are in the *rejection region*, so called because being in the rejection region means that the null hypothesis is rejected.
- The main function for finding cutoffs is the *quantile function*, which takes as its input a probability. (For 1-sided tests, the input is usually $1 - \alpha$; for 2-sided test, the input is usually $1 - \alpha/2$.)
- The quantile function is the inverse of the cdf.
- Examples of quantile functions in SAS include `probit`, `cinv`, `tinvt`, and `finvt`.
- *Example*

- Suppose you are running an F -test at level 0.05 with 3 and 35 degrees of freedom.

- The cutoff for this test is 2.874 (the 95% quantile of an F distribution with 3 and 35 degrees of freedom); thus, F -ratios greater than 2.874 will result in the null hypothesis being rejected.