STAT 5000

Statistical Methods I

Week 2

Fall 2024

Dr. Danica Ommen

Unit 1

Observational Studies

Observational Studies

- In some cases, the treatments cannot be assigned to experimental units by some rule.
  - For example, study of the effects of smoking on cancer with humans as the experimental units
  - Neither ethical nor possible
- We can still gather data by observing some members of the target population as they naturally exist.
  - Census: Observe all members of population
  - Haphazard (convenience) sample
  - Representative random sample
- This type of study is called observational study and is not an experiment.

Observational Studies

- We can still analyze the data from observation studies, but
  - Only about associations, cannot prove causation
  - Only using representative random samples

Simple Random Sampling

Without Replacement: every subset of n unique units has the same probability of being selected (more typical)

With Replacement: on each draw every member of the population has the same chance of being selected and the selected unit is put back into the population before the next unit is selected (some units may be selected more than once)

Simple Random Sampling

- For large populations and small sample sizes:
  - Simple random sample without replacement is similar to simple random sample with replacement.
  - Selection of sample unit changes probability of selection of another sample unit by a very small amount.
  - Treat two sampling schemes in the same way.

Sampling Schemes

- Only consider simple random samples, but there are many other sampling schemes that produce representative samples (Stat 521: Survey Sampling)
- The sampling procedure dictates the method of analysis
- Can make predictions and inferences about associations
- Causal inferences are not justified

Problems Beyond Sampling

- Non-response bias
  - Sampling units do not respond to survey
  - Low response rate leads to convenience sample
  - Methods for increasing response rate - initial introduction, incentives, multiple reminders
- Response bias
  - Non-truthful responses to survey questions
  - Faulty memory, lack of understanding of questions, omissions
- Wording of questions
  - Poor wording - confusing
  - Leading questions

Key Components

- Clear statement of research question(s) and objective(s)
- Identification of target population(s)
- Identification of sampling units
  - Members of the population who provide the measured response
- Measurable characteristics of sampling units (factors)
  - Features of the sampling units
  - Analyze associations with a specified outcome
- Specification of the sampling procedure dictates methods of analysis and restricts types of inference

Types of Observational Studies

Retrospective: potential effects of smoking on lung cancer

- Simple random sample of patients diagnosed with lung cancer at specific set of hospitals
- Independent simple random sample of non-lung cancer patients from same hospitals
- Compare smoking histories for the two samples

Prosepctive: nesting success of pheasants

- Random sample of N locations
- Find nests and implant transmitters in chicks
- Relate survival probability to features of the surrounding habitat

Example: Nurse Health Study

- About 10,000 nurses volunteered to enroll (females who were 20-30 years old at the start)
- Food intake diaries
- Examine association between fat intake and incidence of heart disease
- No control of other factors that might affect incidence of heart disease (genetics, exercise, weight, stress ...)
- Useful information on associations, but cause and effect inferences can not be justified

Example: Fluoridation of Water Supplies and Cancer

- Data collection
  - 10 largest US cities with water fluoridated starting 1951-1956
  - 10 largest US cities that were not fluoridated by 1969
- Cancer deaths per 100,000 population

	1950	1970	change
Fluoridated	180	217	+37
Non-Fluoridated	178	197	+19

- Effect of fluoridation?
  - Only if assume no other difference between cities.

Example: Fluoridation of Water Supplies and Cancer

- “Fluoridation of Water Supplies and Cancer-A Possible Association?" by Oldham and Newell (1997), *Applied Statistics*
- Oldham and Newell’s analysis showed that the two groups of 10 cities differed in their age-sex-race structure in 1950.
- By 1970, the two sets of cities differed much more in their demographic structure than they had in 1950.
- When these demographic changes are taken into account, Oldham and Newell’s analysis showed that “excess” cancers increased 1% in fluoridated cities, 4% in non-fluoridated cities.

Example: Fluoridation of Water Supplies and Cancer

- Data analysis may be based on the sampling distributions of summary statistics, such as sample means
- Data analysis may be based on assumptions about population distributions
  - We will mostly consider analyses for data sampled from populations with normal distributions
  - There are many other distributions, such as the Weibull distribution for survival times
- Central Limit Theorem: The sum (or average) of a large number of independent observations is approximately distributed as a normal random variable

Unit 1

Model-Based Inference:

Distributions

The Normal

A random variable Y with density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

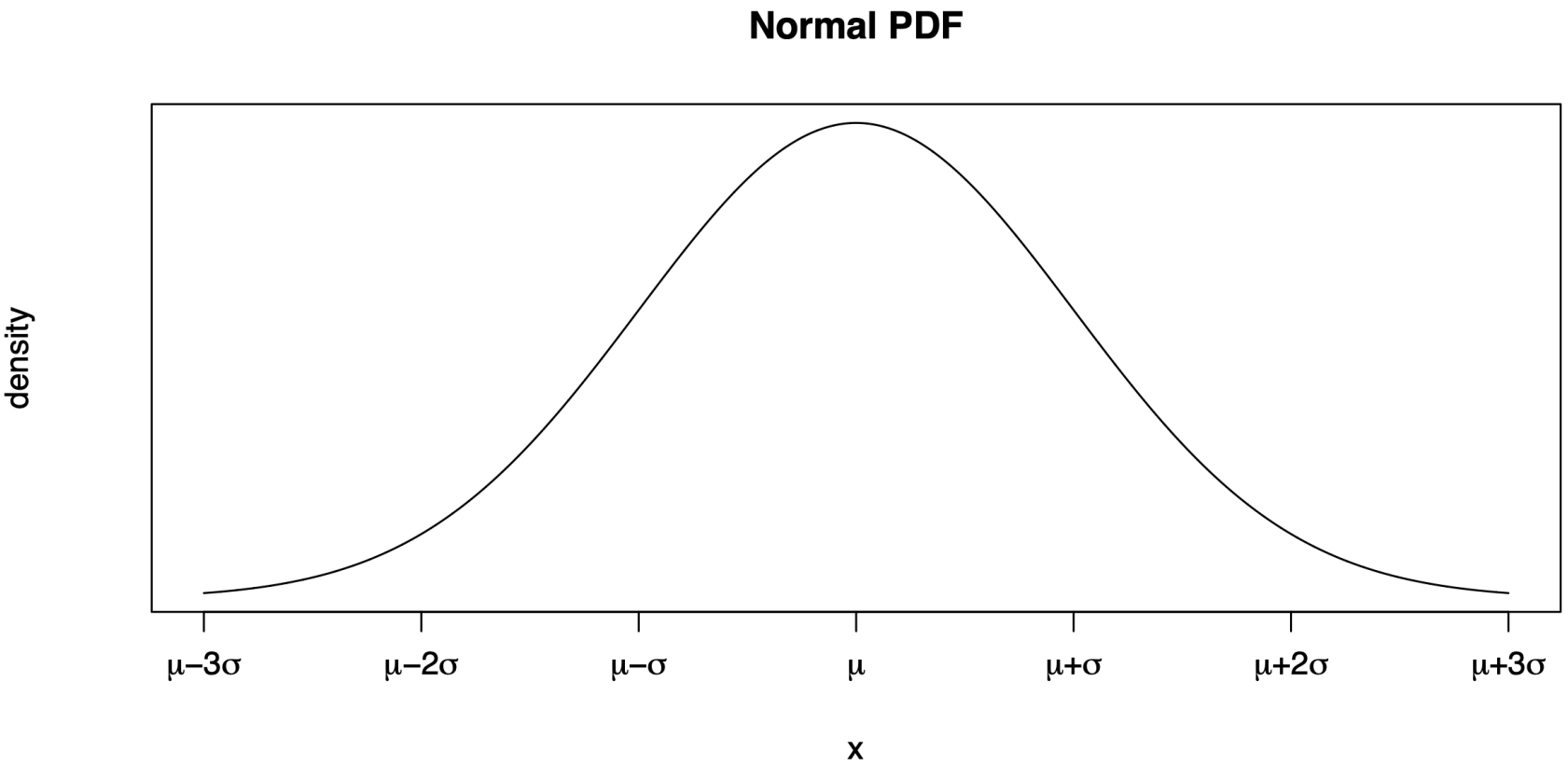
is said to have a normal (*Gaussian*) distribution with

$$\text{Mean} \equiv E(Y) = \mu \qquad \text{and} \qquad \text{Variance} \equiv Var(Y) = \sigma^2$$

The standard deviation is $\sigma = \sqrt{Var(Y)}$

We will use the notation $Y \sim N(\mu, \sigma^2)$

The Normal Density



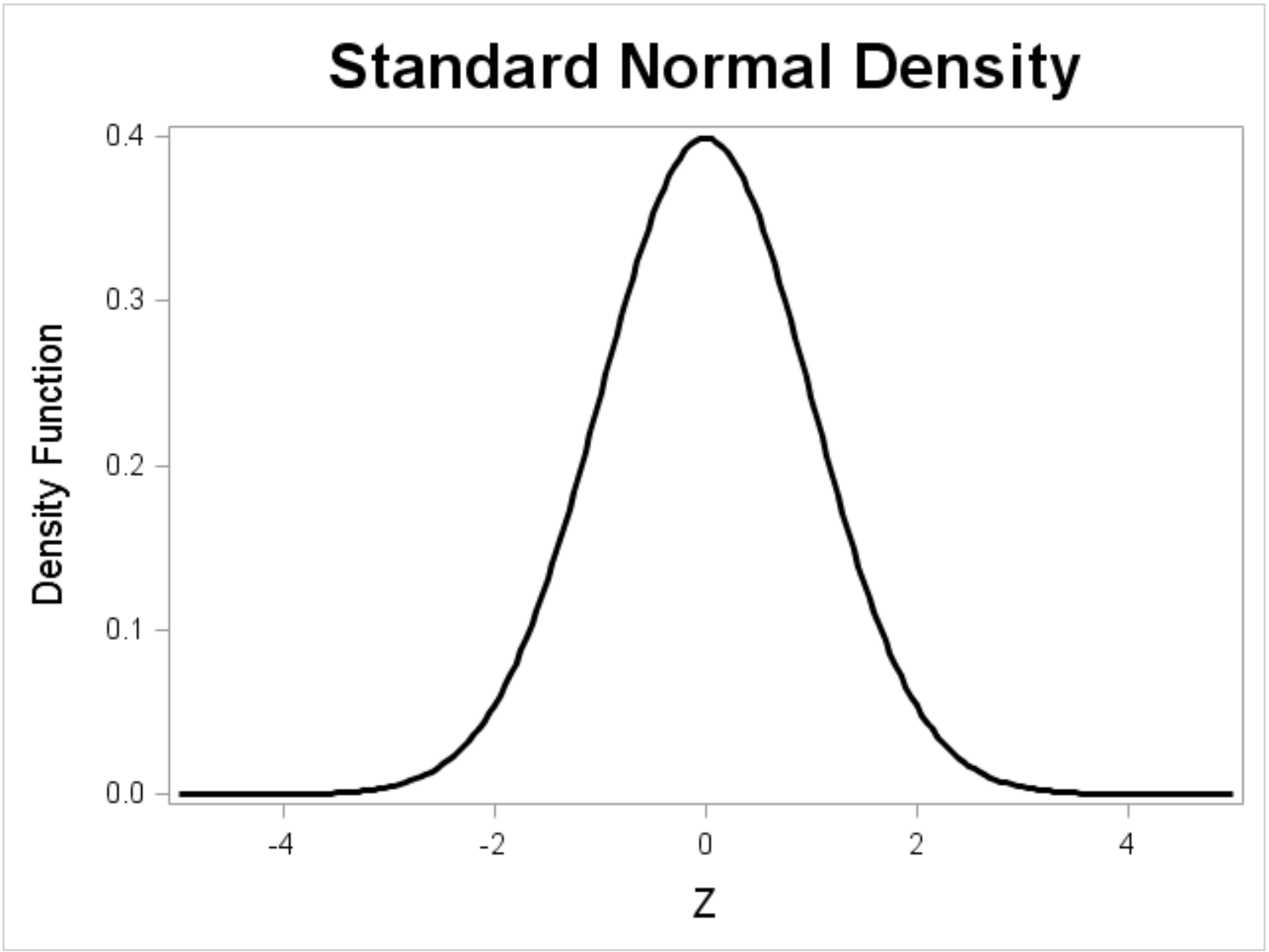
The *Standard* Normal

Suppose Z is a random variable with a normal distribution where $E(Z) = 0$ and $Var(Z) = 1$, i.e.,

$$Z \sim N(0, 1),$$

then Z has a standard normal distribution.

- If $Z \sim N(0, 1)$ then $Y = (\sigma Z + \mu) \sim N(\mu,\ \sigma^2)$
- If $Y \sim N(\mu,\ \sigma^2)$ then $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$


Standard Normal Density
Density Function
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
Z

Linear Combinations

If Y_1 is a random variable with expectation μ_1 and variance σ_1^2 and Y_2 is a random variable with expectation μ_2 and variance σ_2^2, then

- $E(Y_1 + Y_2) = \mu_1 + \mu_2$
- $E(aY_1 + bY_2 + c) = a\mu_1 + b\mu_2 + c$
- $Var(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2$ if Y_1 and Y_2 are independent
- $Var(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2$ if Y_1 and Y_2 are independent
- $Var(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 + 2Cov(Y_1, Y_2)$
- $Var(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2abCov(Y_1, Y_2)$

Useful Definitions

Variance: $Var(Y_1) = \sigma_1^2 = E[(Y_1 - \mu_1)^2]$

Covariance: $Cov(Y_1,\ Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = \rho_{12}\sigma_1\sigma_2$ where ρ_{12} is the correlation between Y_1 and Y_2

The **correlation coefficient** $\rho_{12} = \dfrac{Cov(Y_1,\ Y_2)}{\sigma_1\sigma_2}$ measures the strength of the linear relationship between Y_1 and Y_2. Note that

- It is unit free
- Always between -1 and 1
- Zero when there is no linear association
- Zero if Y_1 and Y_2 are independent of each other

Linear Combinations and The Normal

If $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$ and Y_1 is independent of Y_2 then

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

and

$$aY_1 + bY_2 + c \sim N(a\mu_1 + b\mu_2 + c,\ a^2\sigma_1^2 + b^2\sigma_2^2)$$

A special case of this result with $a = 1$, $b = -1$, $c = 0$ yields

$$Y_1 - Y_2 \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

Distribution of a Sample Mean

- Suppose $Y_{k1}, \ldots, Y_{kn_k}$ denotes a simple random sample of n_k observations from a population with population mean μ_k and variance σ_k^2
- $Y_{k1}, \ldots, Y_{kn_k}$ are *iid* random variables, each with mean μ_k and variance σ_k^2
- The sample mean, $\bar{Y}_k = \sum_{j=1}^{n_k} Y_{kj}/n_k$, is a random variable with expectation

$$E(\bar{Y}_k) = E\left(\frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}\right) = \frac{1}{n_k}\sum_{j=1}^{n_k} E\left(Y_{kj}\right) = \frac{1}{n_k}\sum_{j=1}^{n_k} \mu_k = \mu_k$$

Distribution of a Sample Mean

- The variance of the k-th sample mean is

$$\begin{aligned} Var(\bar{Y}_k) &= Var\left(\frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}\right) \\ &= \frac{1}{n_k^2} Var\left(\sum_{j=1}^{n_k} Y_{kj}\right) \\ &= \frac{1}{n_k^2}\sum_{j=1}^{n_k} Var\left(Y_{kj}\right) \\ &= \frac{1}{n_k^2}\sum_{j=1}^{n_k} \sigma_k^2 \\ &= \frac{\sigma_k^2}{n_k} \end{aligned}$$

Distribution of a Sample Mean

- Assuming independent observations from a population with mean μ_k, the sample mean $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$ is the best linear unbiased estimator for μ_k
- If $Y_{k1}, Y_{k2}, \cdots, Y_{kn_k}$ are *iid* $N(\mu_k,\ \sigma_k^2)$ random variables, i.e., a simple random sample from a normal population, then

$$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj} \ \sim\ N\left(\mu_k, \frac{\sigma_k^2}{n_k}\right)$$

- $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{k,j}$ is a random variable (an *estimator*). Use $\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{k,j}$ to denote it *estimate* (observed value).

Distribution for Difference in Two Sample Means

For independent simple random samples from two normal populations

- $Y_{11}, \ldots, Y_{1n_1}$ are *iid* $N(\mu_1, \sigma_1^2)$ random variables
- $Y_{21}, \ldots, Y_{2n_2}$ are *iid* $N(\mu_2, \sigma_2^2)$ random variables

We can derive that:

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- To draw inference on $\mu_1 - \mu_2$, we need to know σ_1^2 and σ_2^2.
- σ_1^2 and σ_2^2 are population parameters, generally unknown.

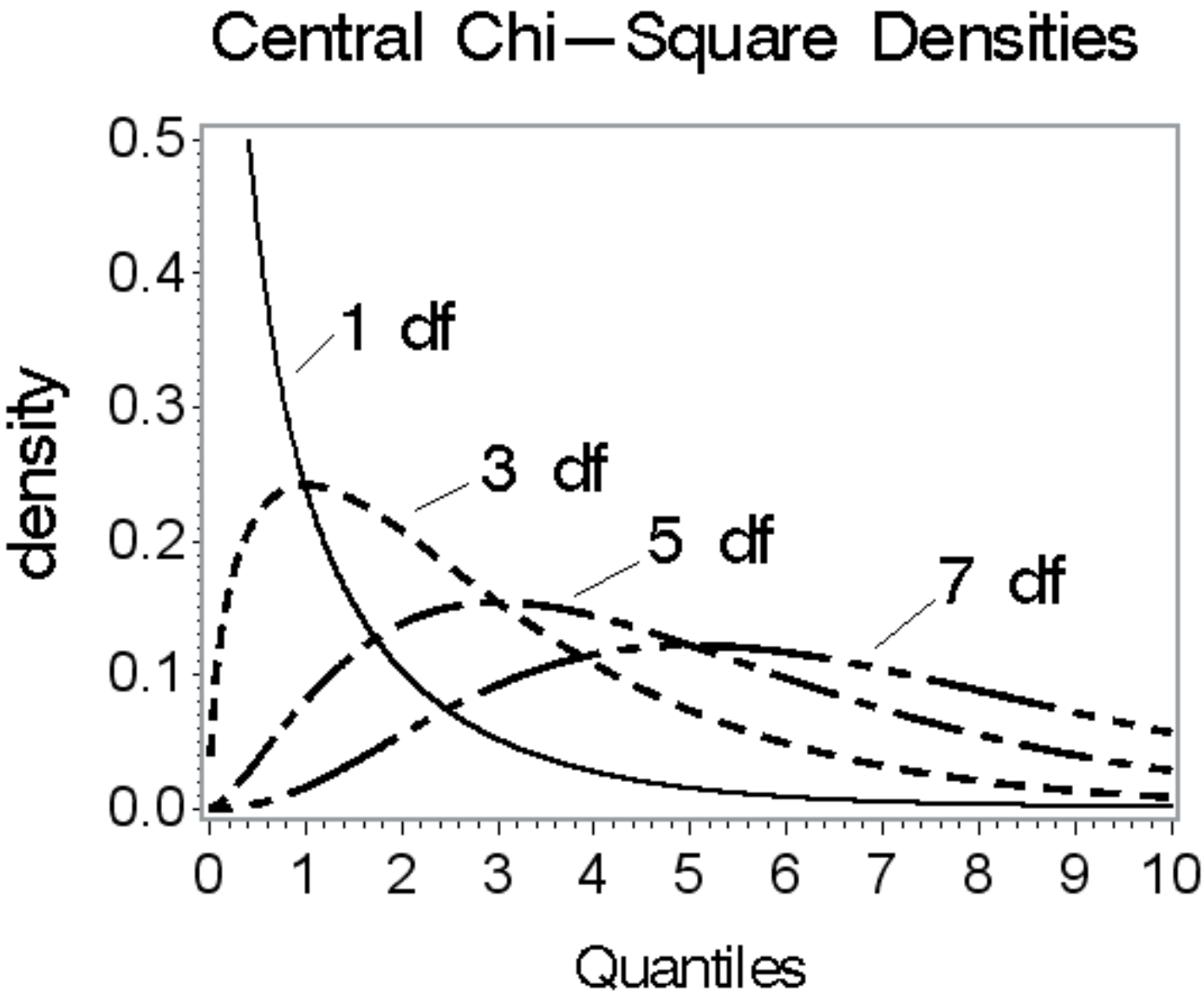
The Central Chi-Square

- Let $Z_i, i = 1, 2, \cdots, n$, be independent standard normal random variables. The distribution of

$$W = \sum_{i=1}^{n} Z_i^2$$

  is called the central chi-square distribution with n degrees of freedom.
- To indicate that a random variable has a central chi-square distribution with ν degrees of freedom, we will use the notation

$$W \sim \chi^2_\nu$$


Central Chi—Square Densities
density
0.5
0.4
0.3
0.2
0.1
0.0
1 df
3 df
5 df
7 df
0 1 2 3 4 5 6 7 8 9 10
Quantiles

Estimation of Variances

For $Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma_1^2)$

- $S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$ is an unbiased estimator of σ_1^2
- $S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$ is an unbiased estimator for σ_2^2
- $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is a pooled estimator for σ^2 where $\sigma_1^2 = \sigma_2^2 = \sigma^2$ indicates homogenous variances

Note: Estimation of Variances

- When $\sigma_1^2 \neq \sigma_2^2$ estimate $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$ as

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

- When $\sigma_1^2 = \sigma_2^2 = \sigma^2$ estimate $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \sigma^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)$ as

$$S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Distribution of a Sample Variance

- It can be shown that

$$\frac{(n_i - 1)S_i^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

  is the sum of the squares of $n_i - 1$ independent standard normal random variables

- Consequently,

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$$

Sum of Independent Chi-Squares

- The sum of two independent central chi-square random variables with ν_1 and ν_2 df, respectively, has a central chi-square distribution with $\nu_1 + \nu_2$ df
- Consequently,

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

has a central chi-square distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom

Note: Degrees of Freedom (df)

- Quantify the amount of information available to estimate a population variance.
- Consider the sample variance $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n_i - 1)$ with $n_i - 1$ df
  - Start with n_i observations
  - Estimate the population mean with $\bar{Y}_i$
  - This imposes one linear restriction $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) = 0$
  - Consequently, the vector of residuals

$$\mathbf{e}_i = (Y_{i1} - \bar{Y}_i,\ Y_{i2} - \bar{Y}_i, \ldots, Y_{in_i} - \bar{Y}_i)^T$$

is an n_i dimensional vector that is restricted to lie in an $n_i - 1$ dimensional linear sub-space

The Student t-Distribution

If $Z \sim N(0,1)$, $W \sim \chi^2_r$, and Z and W are independent random variables, then the random variable

$$T = \frac{Z}{\sqrt{W/r}}$$

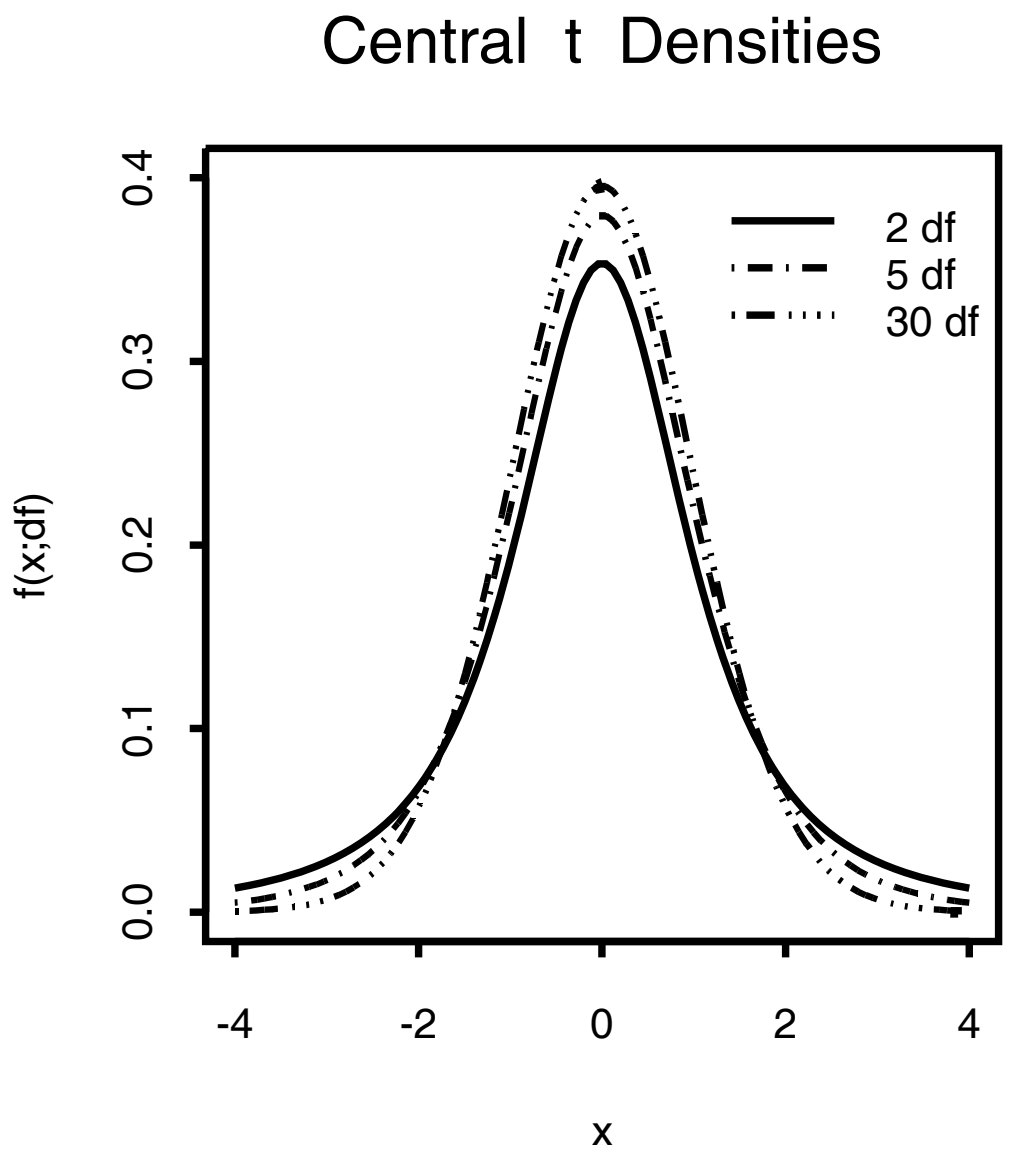
has a central Student t-distribution with r df

To indicate that a random variable has a central t-distribution with r degrees of freedom, we will use the notation

$$T \sim t_r$$

Properties of the Central t-Distribution

- Centered at zero (mean and median are zero)
- Symmetric distribution
- Thicker tails than the standard normal distribution
- Approaches the standard normal distribution as the degrees of freedom become larger
- t_∞ is the standard normal distribution
- 97.5 percentile is around 2 except for small d.f.
  (e.g. 2.571 for 5 d.f., 2.093 for 19 d.f.,
  2.000 for 60 d.f., 1.96 for ∞ d.f.)


Central t Densities
f(x;df)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
2 df
5 df
30 df

Use for Inference:

- $Z = \dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$
- $W = \dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_r$ for $r = n_1 + n_2 - 2$
- Consequently,

$$T = \frac{Z}{\sqrt{W/r}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Inference for Difference in Means with Equal Variances

Assumptions:

- Two independent random samples:
  $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ & $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
- Normality: $Y_{1i} \sim N(\mu_1, \sigma_1^2)$ & $Y_{2j} \sim N(\mu_2, \sigma_2^2)$
- Homogenous Population Variances: $\sigma_1^2 = \sigma_2^2$

Distribution for Inference:

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Questions?

Contact me:

Email: dmommen@iastate.edu
Visit Student Office Hours: Thur @ 10-11 am