Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

September 26, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Use normal probability plots to assess whether a sample is from a normal population.
- Interpret sums of squares, degrees of freedom, and mean squares in a one-factor ANOVA context.
- State the ANOVA partition of the total variation in a data set.
- Carry out a one-factor ANOVA F test for population means $\mu_1, \mu_2, \ldots, \mu_I$.

Notes

Normal Probability Plots

- Two ways to assess the **normality** of data:
  - A ***histogram***. It should be roughly bell-shaped.
  - A ***normal probability plot***. The points should hug the line.

Notes

- Let $X_1, X_2, \ldots, X_n$ be a random sample from *some* distribution.

  Let $\boldsymbol{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}$ denote the same data set, but **sorted** from smallest to largest.

  Thus $X_{(1)}$ is the smallest value in the data set, $X_{(2)}$ is the second smallest, etc.

- The proposition ahead gives the **expected values** of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ **when** the sample is from a **normal** population.

Notes

Proposition

If $X_1, X_2, \ldots, X_n$ are a random sample from a N(μ, σ) distribution, then

$$\begin{aligned} E(X_{(i)}) &\approx 100p_i\text{th percentile of the } N(\mu,\sigma) \text{ distribution} \\ &= \mu + z_i\sigma\,, \end{aligned} \tag{1}$$

where

$$p_i = \frac{i - 0.5}{n}$$

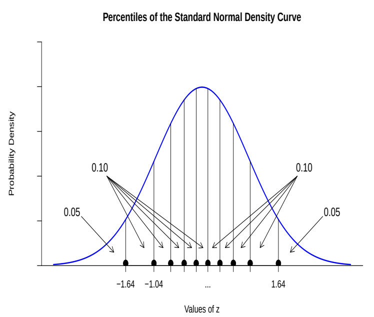
and

$$z_i = \text{The } 100p_i\text{th percentile of the } N(0,1) \text{ distribution.}$$

Notes

- For example, in a sample of size $\boldsymbol{n = 10}$ from a N$(0,\,1)$ distribution, the **expected** sample values are the points shown on the next slide.

Notes

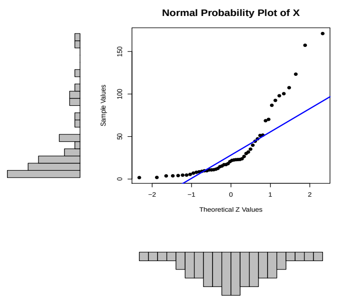


These points are the 5th, 15th, ..., 95th percentiles of the N$(0,\,1)$ distribution:

Notes

- If a sample is from a $N(\mu, \sigma)$ distribution, then
  - The points in a plot of
  $$\left(\mu + z_i\sigma,\ \ X_{(i)}\right),$$
  should fall close to the line $y = x$.
  - The points in a plot of
  $$\left(z_i,\ \ X_{(i)}\right),$$
  should fall close to the line $y = \mu + \sigma x$.
- A ***normal probability plot*** (or ***quantile-quantile plot***) is a plot of the points
$$\left(z_i,\ \ X_{(i)}\right).$$
- **Curved** patterns indicate **non-normality**.

Notes



Notes

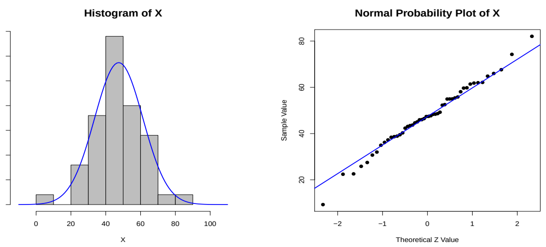


Figure: Histogram of symmetric, approximately normal data (left). Normal probability plot of the same data (right).

Notes

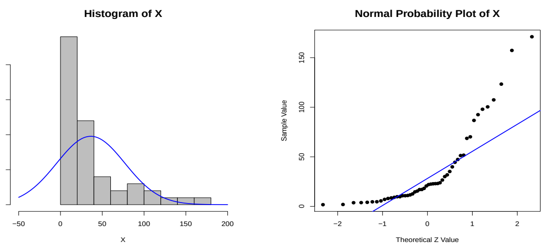


Figure: Histogram of non-normal, right skewed data (left). Normal probability plot of the same data (right).

Notes

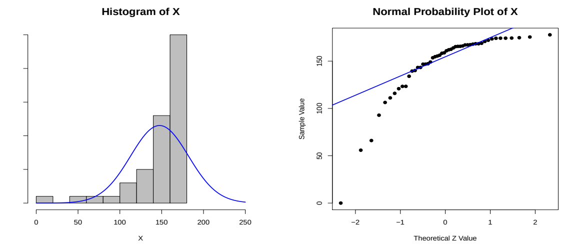


Figure: Histogram of non-normal, left skewed data (left). Normal probability plot of the same data (right).

Notes

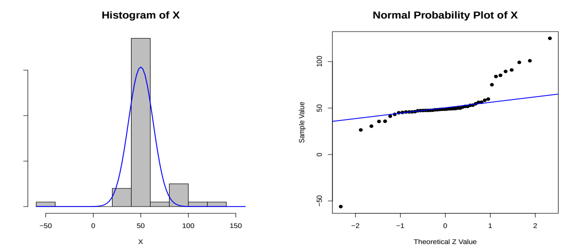


Figure: Histogram of non-normal, ”heavy tailed” data (left). Normal probability plot of the same data (right).

Notes

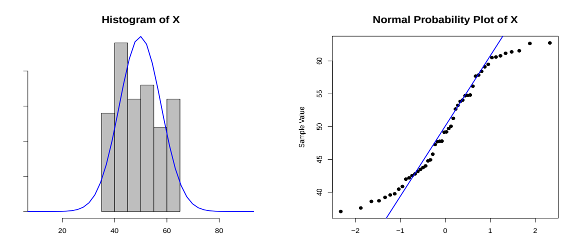


Figure: Histogram of non-normal, ”light tailed” data (left). Normal probability plot of the same data (right).

Notes

One-Factor ANOVA for Population Means $\mu_1, \mu_2, \ldots, \mu_I$

Introduction

- Suppose we have independent random samples from I populations having **possibly different means** but **equal standard deviations**.

  The populations might represent different ***groups*** or they might represent ***treatments*** in an experiment.

  We want to decide if there are any differences among the population means.

Notes

Example

A quality assurance study was carried out to compare **lead measurements** made in water sent to $I = 5$ **laboratories**.

Differences among the five labs' results may signify improperly calibrated equipment or poorly trained technicians.

A vat of wastewater was split into **50** specimens randomized to the labs ($J = 10$ each) for analysis.

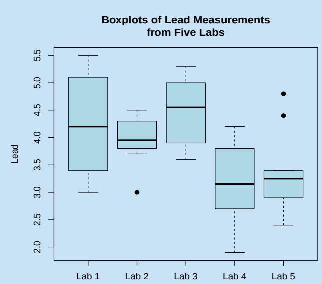
The **lead measurements** (μg/L) and their summary statistics are on the next slide.

Notes

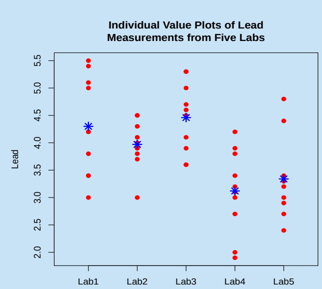
Measured Lead Concentrations

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
3.4	4.5	5.3	3.2	3.3
3.0	3.7	4.7	3.4	2.4
3.4	3.8	3.6	3.1	2.7
5.0	3.9	5.0	3.0	3.2
5.1	4.3	3.6	3.9	3.3
5.5	3.9	4.5	2.0	2.9
5.4	4.1	4.6	1.9	4.4
4.2	4.0	5.3	2.7	3.4
3.8	3.0	3.9	3.8	4.8
4.2	4.5	4.1	4.2	3.0
$\bar{X}_1 = 4.30$	$\bar{X}_2 = 3.97$	$\bar{X}_3 = 4.46$	$\bar{X}_4 = 3.12$	$\bar{X}_5 = 3.34$
$S_1 = 0.904$	$S_2 = 0.440$	$S_3 = 0.642$	$S_4 = 0.764$	$S_5 = 0.737$

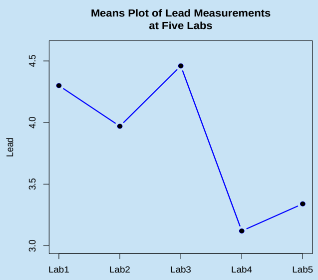
Notes



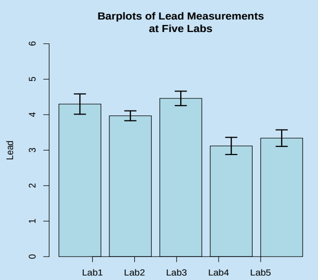
Notes



Notes



Notes



Notes

- Suppose we have random samples, each of size J, from I **populations** $(I \geq 2)$,
- We'll see how to use the samples to decide if there are differences among the **population means** μ_1, μ_2, $\ldots$, and μ_I.

  The appropriate test is called the ***one-factor ANOVA F test***.

Notes

- **Comments**:
  - The sample sizes **don't** all have to be the same. But we'll only look at the equal-sample size case.
  - The data can be **samples** from populations **or** responses to treatments in a **randomized experiment**.

Notes

- The **null hypothesis** is that there are no differences among the population means μ_1, μ_2, ..., μ_I:

Null Hypothesis:

$$H_0 : \mu_1 \;=\; \mu_2 \;=\; \cdots \;=\; \mu_I$$

Notes

- The **alternative hypothesis** is that there's *at least one difference* among the set of means:

Alternative Hypothesis: The alternative hypothesis will be

H_a : At least two of the μ_i's are different

Notes

- **Notation**:

$\boldsymbol{I}$ = The number of treatment groups
$\boldsymbol{J}$ = The common sample size for the I groups
$\boldsymbol{X_{ij}}$ = The jth observation in the ith treatment group
$\boldsymbol{\bar{X}_{i\cdot}}$ = The sample mean for the ith treatment group
$\boldsymbol{S_i}$ = The sample standard deviation for the ith treatment group
$\boldsymbol{\bar{X}_{\cdot\cdot}}$ = The ***grand mean*** of all IJ observations

Note:

$$\bar{X}_{\cdot\cdot} \;=\; \frac{1}{I}\sum_{i=1}^{I} \bar{X}_{i\cdot}$$

(when the sample sizes are all the same).

Notes

Sums of Squares and the ANOVA Partition

- We can ***partition*** the **total variation** in the data into two parts:
  - One reflecting variation **between** the treatment groups.
  - The other reflecting variation **within** the groups.

The ***ANOVA F test*** is based on the amount of **between**-groups variation relative to the amount of **within**-groups variation.

Notes

- The **partition** will involve the following ***sums of squares***:
  - **SST** is the ***total sum of squares***, defined as

$$\text{SST} \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{\cdot\cdot})^2,$$

which measures the **total** variation in the X_{ij}'s.

Notes

- (cont'd):
  - **SSTr** is the ***treatment sum of squares***, defined as

$$\text{SSTr} \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \;=\; J\sum_{i=1}^{I}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2,$$

which measures variation **between** the treatment group means due to both **treatment effects** and **random error**.

Notes

- (cont'd):
  - **SSE** is the ***error sum of squares***, defined as

$$\text{SSE} \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot})^2,$$

which measures variation of the X_{ij}'s **within** treatment groups due to **random error**.

Notes

Proposition

ANOVA Partition of the Total Variation: It can be shown that

$$\text{SST} \;=\; \text{SSTr} \;+\; \text{SSE}.$$

Notes

- The **ANOVA partition** holds because we can write:

$$X_{ij} - \bar{X}_{\cdot\cdot} \quad = \quad \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \quad + \quad X_{ij} - \bar{X}_{i\cdot}$$

Upon squaring both sides and then summing over all i and j, the ”cross product” terms on the right side sum to zero, and we get

$$\sum_i \sum_j (X_{ij} - \bar{X}_{\cdot\cdot})^2 \;=\; \sum_i \sum_j (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \;+\; \sum_i \sum_j (X_{ij} - \bar{X}_{i\cdot})^2 \,,$$

which is the **ANOVA partition**.

Notes

Example

For the data on lead measurements at five labs, software gives

$$\begin{aligned} \text{SST} &= 36.758 \\ \text{SSTr} &= 13.813 \\ \text{SSE} &= 22.945 \end{aligned}$$

The **ANOVA partition** holds:

$$\underset{\substack{\Uparrow \\ \text{Total} \\ \text{variation}}}{36.758} \quad = \quad \underset{\substack{\Uparrow \\ \text{Between} \\ \text{groups} \\ \text{variation}}}{13.813} \quad + \quad \underset{\substack{\Uparrow \\ \text{Within} \\ \text{groups} \\ \text{variation}}}{22.945}$$

Notes

Degrees of Freedom

- Each sum of squares has an associated ***degrees of freedom*** (or ***df***).

The **df** for a sum of squares is determined by how many deviations, among those used to compute the sum of squares, are **”free to vary”** (**unconstrained**).

Notes

Degrees of Freedom:

SST	has	$IJ - 1$ **df**
SSTr	has	$I - 1$ **df**
SSE	has	$I(J-1) = IJ - I$ **df**

Notes

- To see why:
  - The IJ deviations $X_{ij} - \bar{X}_{\cdot\cdot}$ used to compute **SST** are subject to the **one constraint** that they **sum to zero**, i.e.

$$\sum_i \sum_j (X_{ij} - \bar{X}_{\cdot\cdot}) = 0,$$

so only $\boldsymbol{IJ-1}$ of them are "free to vary" (i.e. any $IJ-1$ of them determines the remaining one).

Notes

- (cont'd):
  - The I deviations $\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}$ used to compute **SSTr** are subject to the **one constraint** that that they **sum to zero**, i.e.

$$\sum_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) = 0,$$

so only $\boldsymbol{I-1}$ of the deviations are "free to vary" (i.e. any $I-1$ of them determines the remaining one).

Notes

- (cont'd):
  - The IJ deviations $X_{ij} - \bar{X}_{i\cdot}$ used to compute **SSE** are subject to the $\boldsymbol{I}$ **constraints** that they **sum to zero *within* each of the $\boldsymbol{I}$ groups**, i.e.

$$\sum_j (X_{ij} - \bar{X}_{i\cdot}) = 0 \qquad \text{for each } i = 1, 2, \ldots, I$$

Thus within each of the $\boldsymbol{I}$ samples, only $\boldsymbol{J-1}$ deviations are "free to vary" (i.e. any $J-1$ of them determines the remaining one).

Notes

Additive Property of Degrees of Freedom:

$$\text{df for SST} = \text{df for SSTr} + \text{df for SSE}$$

since

$$IJ - 1 = (I-1) + I(J-1).$$

Notes

Notes

Mean Squares

- The ***ANOVA F test*** is based on the amount of **between**-groups variation relative to the amount of **within**-groups variation.

  But **SSTr** and **SSE** *aren't* directly comparable (they depend in different ways on I and J).

- A ***mean square*** a **sum of squares** divided by its **df**.

  Example: A *sample variance* S^2 is a **mean square**.

Notes

- (cont'd)
  - The ***mean square for treatments***, denoted **MSTr**, is

$$\text{MSTr} = \frac{\text{SSTr}}{I-1}.$$

Notes

- (cont'd)
  - The ***mean squared error***, denoted **MSE**, is

$$\text{MSE} = \frac{\text{SSE}}{I(J-1)}.$$

It's easy to verify that

$$\text{MSE} = \frac{S_1^2 + S_2^2 + \cdots + S_I^2}{I}$$

(when the sample sizes are all the same).

Thus **MSE** is the **average** (or **pooled**) **sample variance**.

Notes

- **MSTr** and **MSE** *are* directly comparable.

The One-Factor ANOVA F Test

One-Factor ANOVA F Test Statistic:

$$F = \frac{\text{MSTr}}{\text{MSE}}.$$

Notes

- F reflects **between**-groups variation (**MSTr**) relative to **within**-groups variation (**SSE**).
- **MSTr** will be **large** when there's substantial variation in $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{I\cdot}$, which are estimates of the population means $\mu_1, \mu_2, \ldots, \mu_I$.

  It will be **large** when there are **differences** among $\mu_1, \mu_2, \ldots, \mu_I$.

Notes

Large values of F provide **evidence against H_0 in favor of H_a : At least two of the μ_i's are different**.

Notes

- Now suppose the I samples are from $N(\mu_1, \sigma)$, $N(\mu_2, \sigma)$ $\ldots$, $N(\mu_I, \sigma)$ distributions and that they were drawn *independently* of each other.

  Alternatively, the samples could be from **non-normal** populations as long as the common sample size J is **large**.

Notes

Sampling Distribution of the Test Statistic Under H_0: If F is the one-factor ANOVA F test statistic, then when

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

is true,

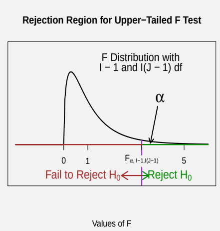
$$F \sim F(I-1,\, I(J-1)).$$

Notes

- The $F(I-1,\, I(J-1))$ curve gives us:
  - The ***rejection region*** as the **extreme largest 100α% of F values**.
  - The ***p-value*** as the **tail area to the right of the observed F value**.

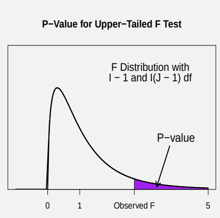
Notes

Rejection Region: The **rejection region** is the **set of F values** in the tail of the $F(I-1, I(J-1))$ curve to the **right of $F_{\alpha,\, I-1,\, I(J-1)}$**:



Notes

P-Value: The **p-value** is the **tail area** under the $F(I-1, I(J-1))$ curve to the **right** of the **observed F**:



Notes

The ANOVA Table

- ANOVA results are summarized in an ***ANOVA table***:

Source of Variation	df	Sum of Squares	Mean Square	f	P-value
Treatment	$I-1$	SSTr	MSTr = SSTr/(I-1)	MSTr/MSE	p
Error	$I(J-1)$	SSE	MSE = SSE/(I(J-1))		
Total	IJ-1	SST			

Notes

Exercise

For lead measurements made at five labs, the **ANOVA table** is:

Source of Variation	df	Sum of Squares	Mean Square	f	P-value
Treatment	4	13.813	3.453	6.77	0.000
Error	45	22.945	0.510		
Total	49	36.758			

Notes

a) Verify that **df for SSTr =** $I-1$, that **df for SSE =** $I(J-1)$, and that **df for SST =** $IJ-1$.

b) Verify that **SST = SSTr + SSE** and that the **df for SST = df for SSTr + df for SSE**.

c) Verify that the **mean squares** are the **sums of squares** divided by their **df**.

d) Verify that the F **statistic** is **MSTr** divided by **MSE**.

Notes

e) State the **hypotheses**.

f) Using $\alpha = 0.05$, is there statistically significant evidence for systematic differences in lead measurements among the five labs?

g) If there are significant differences among the five labs, describe the nature of those differences (using the plots of the data given earlier in these slides).

Notes

- For comparing **two population means** μ_1 and μ_2, the **ANOVA** F **test** and a *two-sided* **pooled two-sample** t **test** are **equivalent**.

  The **square** of the t **statistic** is the F **statistic**, and the **p-values** will be the **same**.

Notes

Example

An example in a previous set of slides presented results of a computer simulation to compare the time (in seconds) to complete a semiconductor manufacturing process using one and two operators.

Here are the summary statistics:

One Operator	Two Operators
$m = 16$	$n = 16$
$\bar{X} = 373.6$	$\bar{Y} = 374.8$
$S_1 = 7.8$	$S_2 = 7.3$

Notes

If we carry out a **(pooled) two-sample** t **test** of

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

we get:

Pooled t **Test Statistic**	**P-Value**
$t = -0.445$	**0.6596**

Notes

If we carry out a **one-factor ANOVA**, we get:

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**f**	**P-value**
Treatment	1	11.3	11.3	0.198	0.6596
Error	30	1710.2	57.0		
Total	31	1721.5			

We see that $t^2 = F$ and the **p-values** for the two tests are the same.

Notes

- In general, the **square** of a t random variable is an F random variable.

Proposition

If
$$T \sim t(\nu)$$
then
$$T^2 \sim F(1, \nu).$$

Notes

Notes

Notes

Notes