

✧ RNR 613 — Descriptive Statistics

Describing populations with their sample distributions

Examine distributions graphically

- *Histograms, Step-and-Leaf Plots*

Frequency histograms: plots of the observed values of some variable (y) versus the number of times (*frequency*) each value was recorded. Generate with a *stem-and-leaf plot*. First determine the interval for the *step*, then place the *leaves* on the *tree*.

E.g., consider a set of grades for an ecology course which ranged from 40-95. Might choose 50, 60, 70, 80, 90 as the step. First 10 grades (79, 87, 76, 72, 63, 89, 52, 93, 69, 73) yield this stem-and-leaf plot:

```

5 | 2
6 | 3 9
7 | 9 6 2 3
8 | 7 9
9 | 3

```

- *Box-Plots* — We will discuss later...

Describe distributions quantitatively

With a series of descriptive statistics, including a series of measurements called *moments* $\sum (y_i - \mu)^p / N$, where p represents the p^{th} moment about the mean.

Measures of Central Tendency or Location

Sample values often group near the middle of the range of observed values. In these cases, it can be useful to describe the center of a distribution with *measures of central tendency* (sometimes called *measures of location*). These describe concentrations of values in a sample.

- **Arithmetic Mean** (first moment)

Also called *average*. If each measurement is denoted as y_i , where i ranges from 1 to N , and N is the

number of observations in a population, then the *population mean* (μ) is:
$$\mu = \frac{\sum_{i=1}^N y_i}{N}$$

The population mean is a parameter estimated with the *sample mean* or *sample average*, denoted by \bar{y} .

The *sample average* is an unbiased estimator of the population mean [$E(\bar{y}) = \mu$] and is calculated similarly

except that N is replaced by *sample size*, n :
$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

For the grade data, $\bar{y} = (79+87+76+72+63+89+52+93+69+73)/10 = 75.3$

The mean or average describes the *center of mass* or “fulcrum” on which a histogram would balance horizontally.

In addition to the *arithmetic average*, other “averages” are useful, such as a *weighted average* (an average

of averages):
$$\bar{y}_w = \frac{\sum_{i=1}^n w_i \bar{y}_i}{\sum_{i=1}^n w_i}$$
 where w_i is the *weighting factor* for each value.

Using sample size as the weight, for example:

\bar{y}_i	n_i	$\bar{y}_i \times n_i$
3.85	12	46.20
5.21	25	130.25
4.70	8	37.60
	45	214.05

weighted mean = $214.05/45 = 4.76$

arithmetic mean = $(3.85 + 5.21 + 4.70)/3 = 4.59$

• Median

The middle measurement in an ordered (sorted, ranked) sample. There are an equal number of observations that are greater than and less than the *median*.

For the (sorted) grade data: 52 63 69 72 73 76 79 87 89 93; sample median = $(73 + 76)/2 = 74.5$

The median incorporate no information about other data in a sample, so is not sensitive to extreme values in the sample. In general, statistics that are insensitive to "outliers," such as the median, are called *resistant*.

• Quantiles

Extensions of the median, *quantiles* dissect distributions into some number of equal portions.

E.g., dividing a sample in 4 equal portions yields *quartiles*, each of which contain 25% of the sample data;

25% of the ranked observations are less than the first quartile,
 25% lie between the first and second quartiles,
 25% lie between the second and third quartiles,
 25% are greater than the third quartile.

The difference between the 25% and 75% quartiles is the *interquartile range*.

• Mode

The most frequently occurring measurement in a sample. The mode represent points of concentration in a distribution, and only makes sense for *unimodal* distributions.

Interrelationships between mean, median, mode

- If sample data form a *Normal* distribution, then $mean = median = mode$; all of which located in the exact center of the distribution.
- If the distribution is *asymmetric or skewed*, then the relationship of these 3 measures depends on whether the *skewness* is positive or negative.
- If a distribution is *skewed positively* (has a long tail to the right), then $mode < median < mean$.
- If the distribution is *skewed negatively* (has a long tail to the left), then $mean < median < mode$.

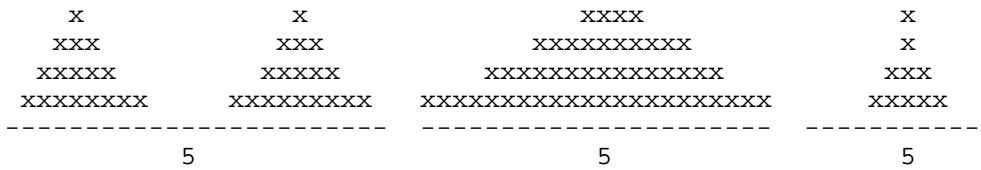
• **Skewness** (third moment) — measures asymmetry of a distribution.

• **Kurtosis** (fourth moment) — measures relative *shape* of the *middle and tails*: *peakedness*.

Measures of Dispersion or Variation

Central tendency is only one attribute of a distribution. Two distributions with the same *sample mean*, for

example, may be very different.



The following attributes describe *dispersion or variability*, i.e., how “spread out” observations are about a distribution’s center.

- **Range**

Sample range is the difference between the largest and smallest observations in a sample, which is a poor estimate of the *population range*, as a sample is unlikely to contain the lowest and highest values in the population. The *sample range* does not take into account any other observations in a sample.

- **Mean Deviation**

Deviation is a measure of the “distance” each observation is from the mean: $|y_i - \bar{y}|$ and is our first *parametric* statistic (*definition*: one value [a constant] used as a reference for determining other values).

Sample mean deviation is the sum of *absolute deviation* for all observations divided by n :

$$\text{sample mean deviation} = \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n}$$

- **Variance** (second moment)

An alternate way of eliminating negative values from the sum of deviations is to square each absolute deviation rather than take its absolute value: $\sum (y_i - \bar{y})^2$

This important quantity is the *sum of squared deviations* or *sum of squares (SS)*.

The sum of squared deviations from the mean is smaller than the sum of squared deviations from any other quantity (such as the median). This principle is called *Least Squares*.

The average SS for all observations yields a measure of the average distance an observation would deviate

from the mean: $\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$. This quantity is (almost) *sample variance*.

$$\text{Population variance is: } \sigma^2 = \frac{\sum_{i=1}^N (y_i - \mu)^2}{N}; \text{ Sample variance is: } s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

Units for variance are the original units squared. If all observations in the sample are equal, then $s^2 = 0$. Because variance is a squared quantity, as sample variability increases, variance increases to the square. For the grade data, $SS = 7770.1$; $s^2 = SS / (10 - 1) = 7770.1/9 = 863.24$.

- **Standard Deviation (SD)**

The square root of variance yields a quantity in units of the original measurements, the *population standard*

$$\text{deviation: } \sigma = \sqrt{\frac{\sum_{i=1}^N (y_i - \mu)^2}{N}}$$

Similarly, *sample standard deviation* is (divided by $n - 1$):
$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

SD is also called *root mean square* in ANOVA and regression. For the grade data, $SD = \sqrt{863.2} = 29.4$.

• Coefficient of Variation (CV)

CV expresses sample standard deviation relative to sample mean. *Sample CV* = s/\bar{y} or $s/\bar{y} * 100\%$. This is useful when comparing variation between groups on different scales of measurement.

For the grade data, $CV = 29.4/75.3 * 100 = 39.4$.

• Standard Error

Standard error is a characteristic of a *statistic* rather than a sample; it estimates the standard deviation of a statistic's *sampling distribution*.

It is a measure of precision and represents the best "guess" of the size of the difference between a statistic's *estimate* and its *true value*.

Standard error of the mean (SEM) is denoted as $\sigma_{\bar{y}}$ and estimated as:
$$s_{\bar{y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

SEM describes how precisely the sample mean was measured; this proves useful in hypothesis testing.

Central Limit Theorem

An important characteristic of the sampling distribution of means (formed from calculating means of multiple samples collected from a single population): even if the distribution of the population sampled is not normal, the distribution of the *sample means* will be normal, especially as sample size increases.

Characteristics of moments about the mean:

1. $\sum (y_i - \mu)^1/N$: always equal to zero; the absolute value of this quantity is *mean deviation*;
2. $\sum (y_i - \mu)^2/N$: *variance* (also called *mean squared deviation* or *mean square*); square root of variance is *standard deviation* (also called *root mean square*); numerator of variance, $\sum (y_i - \bar{y})^2$, is called *sum of squared deviations*, or simply, *sum of squares (SS)*;
3. $\sum (y_i - \mu)^3/N$: *skewness* describes symmetry of a distribution;
4. $\sum (y_i - \mu)^4/N$: *kurtosis* describes peakedness of a distribution; (heavy right tail = +; heavy left tail = -)

Symbols for parameters and their associated estimators.

Population Parameter	Symbol	Sample Estimator	Symbol	Measures
Mean	μ	Average	\bar{Y}	Central Tendency of Distribution
Variance	σ^2	Variance	s^2	Deviation of Distribution
Standard Deviation	σ	Standard Deviation	s	Deviation of Distribution
Standard Error of the Mean	$\sigma_{\bar{y}}$	Standard Error	$s_{\bar{y}}$	Precision of Mean

✧ RNR 613 — Estimation, Testing, Design, Inference

Statistical tools used depend on research *Questions* or *Goals*

- *Goal*: estimate the value of a population parameter (e.g., population mean, μ , or treatment effect, δ) with sample data:

→ *Parameter Estimation*

e.g., What are the plausible values for the mean height of Saguaro cactus on UA campus?

- *Goal*: test a specific null hypothesis about a parameter:

→ *Hypothesis Testing / Inferential Statistics*

e.g., Conclude whether or not supplemental water increases germination rate of Saguaros (*or identically*: Determine if it is plausible that the effect of supplemental water on germination was zero).

Tools for *estimation* and *testing* are related closely.

For the hypothesis test above, a related and meaningful question might be to determine the plausible values by which germination rates increased by addition of supplemental water...

Why use statistical tools for parameter estimation?

Goal: Assign some degree of confidence to how precisely the sample estimates the population parameter.

Tool: Confidence Intervals.

Why use statistical tools for hypothesis testing?

Goal: Determine if a population parameter differs from some hypothesized value.

E.g., compare means of some parameter between two populations, perhaps body mass of male and female bald eagles.

Tool: Test the null hypothesis that $\mu_{\text{males}} = \mu_{\text{females}}$ or identically, $\mu_{\text{males}} - \mu_{\text{females}} = 0$, where we wish to determine if the parameter of interest (difference between mean body mass of males and females) differs from a hypothesized value (here, 0).

The precision with which we estimate parameters affects both the strength (power) of the hypothesis test (whether or not to reject the null hypothesis) and the width of the confidence interval around the estimate.

Experimental Units

In an experiment, an *experimental unit* is the smallest independent unit to which a treatment can be assigned.

Identifying the experimental unit (or sampling unit in some cases) is critical to performing the appropriate analysis.

Inference

Drawing conclusions about the state of some set of things based on knowledge about another set of things. We often want to infer characteristics of a population based on sample data.

Inference can be either *inductive* or *deductive*.

Inference is inductive if conclusions about a general case are drawn from information about a specific case.

E.g., driving home you observe cars on a freeway overpass in gridlock. From that “sample,” you make an *inductive inference* about the condition of the freeway.

Inference is deductive if conclusions about a specific case are drawn from information about a more general case.

E.g., driving home you hear a traffic report stating that the freeway is jammed. You have made a *deductive inference* if you conclude that the portion of the freeway on which you must travel, in particular, is jammed based on information about the freeway as a whole.

Statistical reasoning is inductive

Statistical methods use laws of probability to understand relationships between what we observe and the conditions that lead to what we observe, as expressed by probabilities of various events.

A *statistical inference* is an inference justified by a probability model linking data to the broader context.

The type (and strength) of inference drawn depends crucially on the research/experimental design, which fall into two distinct categories: *randomized experiment* and *observational study*.

In a *randomized experiment*, the researcher controls the assignment of experimental units to treatment groups and uses a chance mechanism to make the assignment.

In an *observational study*, the group status of the subjects is beyond the control of the researcher.

Two forms of inference — *causal inference* and *inference to populations* — can only be justified when randomization mechanisms (chance) are employed by the researcher.

Causal Inference

Causal = Establish *Cause-and-Effect*.

Statistical inferences of cause-and-effect relationships can be drawn only from *randomized experiments*, not from *observational studies*.

Randomization “mixes up” experimental units with different and possibly relevant features between treatment groups, reducing the chance of *confounding*.

Confounded variables are related to both:

- (1) group membership
- (2) the outcome measured in the experiment (response).

Their presence makes it difficult to establish the outcome as a direct consequence of group membership.

Inference to Populations

This distinction relates to how units (subjects) are selected from a larger population.

In a *random sampling study*, units are selected by the researcher from a well-defined population. All units in the population have the same chance of being selected and selection is through a chance mechanism.

In a *self-selection study*, units are predetermined, which precludes the units from representing a broader population.

*** Inferences to *populations* can be drawn from random sampling studies, but not otherwise.

In Summary

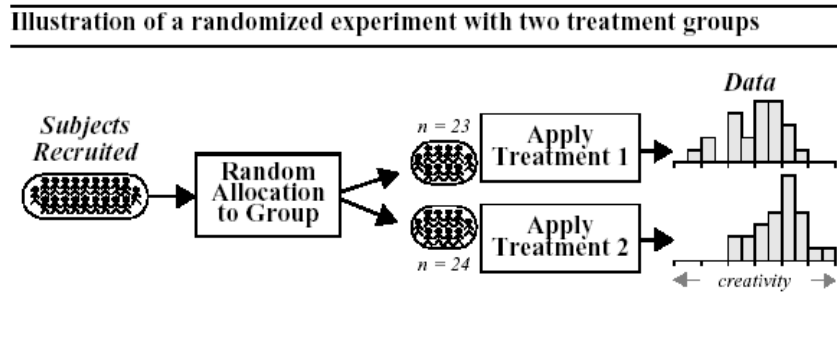
Two levels of inference for all studies:

- *Selection of Units:*
At random: Inferences to population can be drawn.
Not at random: Inferences to population cannot be drawn.
- *Allocation of Units:*
By randomization: Causal inferences can be drawn.
Not by randomization: Causal inferences cannot be drawn.

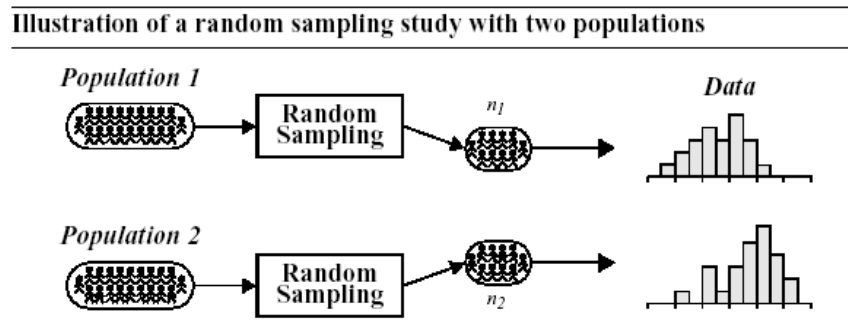
✿ RNR 613 — Uncertainty and Probability Models

- Occurrence of events are uncertain.
- Parameters are unknown; we can only estimate them.
- Models to describe data are never perfect; statistics help us to assess if a model seems reasonable compared to alternative models given the data.

Uncertainty in Randomized Experiments



Uncertainty in Observational Studies



Probability Model for Randomized Experiments

Additive Treatment Effect

An *additive* treatment effect describes the situation where the response variable measured changes in an additive (+) way after treatment is applied.

Let Y^* denote the value of a response variable measured from subjects after treatment has been applied (Y is the value of the response before treatment was applied):

$$Y^* = Y + \delta$$

The treatment effect, δ , is an unknown constant (parameter) that describes a feature in the model used for answering questions of interest.

- *A Test for Treatment Effect*

The question of interest is often whether or not a treatment effect is “meaningful,” which we assess with a *probability value* or *P-value*.

*** *P-values* express the uncertainty of the inferences made, and represent the probability of obtaining a value of the *test statistic* at least as extreme as the value generated from the sample data. It is sometimes viewed as the degree of evidence suggesting that the null hypothesis is true.

A *test statistic* is calculated from the data. The appropriate test statistic depends on the “context” and the analysis performed.

- *Null and Alternative Hypotheses*

The question of interest “Is there a non-zero treatment effect?” is cast as a question about whether δ is or is not zero.

If $\delta = 0$ then the answer to the question is **no**;
If $\delta \neq 0$ then the answer to the question is **yes**.

The statement that $\delta = 0$ is the *null hypothesis* (denoted H_0);
The statement that $\delta \neq 0$ is the *alternative hypothesis* (H_A).

The way the alternative hypothesis is cast is important because it describes how to summarize the sample data to answer the question of interest.

If $\delta \neq 0$, we estimate the value of δ using the difference between the response variables measured for each treatment group.

With two groups *treatment* and *control*, $\hat{\delta} = \bar{Y}_T - \bar{Y}_C$

E.g., if $\bar{Y}_T = 15.74$ and $\bar{Y}_C = 19.88$, then $\hat{\delta} = 4.14$, which is the appropriate test statistic for this question (which depends on how the alternative hypothesis is stated).

- *Randomization Distribution of the Test Statistic*

If a treatment had no effect, subjects responses would be the same regardless of group assignment.

We could determine the value of the test statistic if we “pretend” that group assignment (the original randomization) had turned out differently. In fact, we could calculate the value of the test statistic for every possible randomization.

Stated another way, we could take all the responses measured, then assign group membership in all possible combinations.

A histogram of these values is a *randomization distribution*.

Uncertainty in Observational Studies

Is handled similarly to that of Randomized Experiments except that in sampling studies, questions focus on how features of the population differ among pre-existing groups of interest.

In many cases, differences among groups can be described by the difference between their means, so that $\mu_2 - \mu_1$ is the parameter of interest.

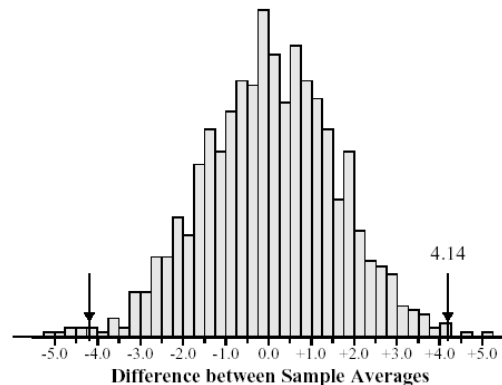
Inference and uncertainty measures are based on $\bar{y}_2 - \bar{y}_1$ which estimates the difference between population means of 2 groups.

A *sampling distribution* for the test statistic $\bar{y}_2 - \bar{y}_1$ is a histogram of all values of the statistic from all possible samples that can be drawn from the two populations.

A *permutation distribution* of this test statistic portrays the collection of differences in averages from all possible group assignments to subjects.

From this distribution we can calculate an exact *P*-value.

A histogram of differences between group averages, from 1,000 randomizations of the creativity study data



✧ RNR 613 — Hypothesis Testing

Statistics are an essential element in the toolkit of scientists; they help us separate signal from noise, anecdotal conjecture from real evidence, and pattern from randomness.

We have placed too much emphasis on the results of hypothesis tests and too little on the scientific information in the data itself.

Hypothesis Testing

Compare the mean of some attribute between groups ($\mu_2 - \mu_1$):

1. What are the likely values that might result from $\mu_1 - \mu_2$,
2. Is zero (or some other value) a likely value to expect as a result from $\mu_1 - \mu_2$?

Answer question (1) with *estimation* and *confidence intervals* and question (2) with *hypothesis tests*.

• The process...

- State a null hypothesis about the population mean, which is often a finding of no effect (H_0). E.g., $H_0: \mu = 3$ m.
- State an *alternate hypothesis*, against which the null will be judged. E.g., $H_a: \mu \neq 3$ m or $H_a: \mu > 3$ m or $H_a: \mu < 3$ m.

Note that null and alternate hypotheses are always complements of each other, can be one- or two-directional.

- Collect sample data and then determine the probability, assuming H_0 to be true, of attaining a \bar{y} at least as far from μ as the \bar{y} observed in the sample using a *test statistic*.
- If we conclude the null hypothesis to be false, *reject the null hypothesis* in favor of the alternative hypothesis.
- If we *do not reject the null hypothesis*, can we conclude the null hypothesis to be true?

No. We can never accept a null hypothesis as true, we can only reject or not reject.

Example. A manufacturer wants a new alarm to sound when the concentration of CO (mg/m³) in air is 10 mg/m³.

The null hypothesis can be stated as $H_0: \mu = 10 \text{ mg/m}^3$
The alternate hypothesis as $H_A: \mu \neq 10 \text{ mg/m}^3$.

He then adds CO and records the concentration at which the alarm sounds 18 times. For these observations:

$$n = 18 \quad \bar{y} = 10.43 \quad s^2 = 1.043 \text{ (mg/m}^3\text{)}^2$$

$$s_{\bar{y}} = \sqrt{(1.043/18)} = 0.24 \text{ mg/m}^3$$

If we assume a *Normal* population with $\mu = 10$ and $\sigma_{\bar{y}} = 0.24$, what is the probability of obtaining a random sample with a mean (\bar{y}) at least as far from 10 as the mean observed, 10.43?

Identically, what is $P(\bar{y} \geq 10.43 \text{ or } \bar{y} \leq 9.57)$?

Put another way: If the null hypothesis is indeed true, samples drawn from this population have $\mu = 10$ and $\sigma_{\bar{y}} = 0.24$ and would look like this: (overhead)

$$Z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}$$

The probability of obtaining a value *at least as extreme as our sample value* is the sum of the area under the curve that is ≥ 10.43 and ≤ 9.57 .

This is the *p-value* that represents the probability of obtaining a value at least as extreme as our sample value given the data.

In this example, the probability can be calculated by computing a Z-statistic, which is the distance between null and alternative hypotheses standardized by SEM.

Stated more simply, how many SEM is \bar{y} from μ if the null hypothesis is actually true?:

$$\text{In the CO example: } Z = \frac{10.43 - 10}{0.24} = 1.79$$

As $Z = 1.79$, so \bar{y} is 1.79 SEM from μ , which based on a Normal distribution yields a *p-value* = 0.0734.

This *p-value* suggests that we would expect to obtain a value at least as extreme as our sample value 7 of 100 times.

This *p-value* represents an area under the Normal curve which we can derive exactly with a computer (see link on course web page) or using a table for an approximation. We still have not interpreted the Z-test statistic and *P-value*, and therefore have not made a decision about the hypothesis test.