Five Number Summary

2	14	28	29	30	32	33	34	40	42	52
Min		Q1			Med			Q3		Max

Minimum, Q1, Median, Q3, Maximum

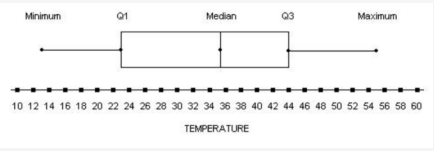
IQR and Outliers

IQR: Q3-Q1

Low Outlier: (Q1)(1.5 IQR)

High Outlier: (Q3)(1.5 IQR)

Boxplot



Histograms

Regular: equal spacing, used to determine shape

Relative: divide # by n to get %

Cumulative: ≤, last bar = n, never for shape

Cumulative: percentage based, 50% is median

relative location= (# of values below)/n

Transformations

Linear: affects center not spread

Adds to sample mean (x), M, Q1, Q3, IQR

Multiplication: affects center and spread

Multiplies to mean (x), M, Q1, Q3, IQR, and sigma

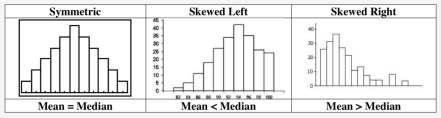
"Describe the distribution"

Center: Mean or median

Spread; Standard deviation, IQR or range

Shape: Symmetric or skewed

Shape



Density Curves

-Always on or above x-axis
-Area = 1
-Has mean and median
-Infinite tails
-Singular point has NO AREA

Median: equal areas point (point that divides curve in half)
Mean: balance point (where curve would balance if solid)
Symmetric curve: Mean = Median

"Describe the relationship" (Scatter plot)

Direction: +/- or neither (analyzed from L to R)

Form: linear or not

Strength: how correlated (see below)

Very strong: 100-91
Strong: 90-85
Moderately strong: 84-75
Moderately weak: 74-70
Weak: 70 and below

Residual Plots

Exhibits randomness, then a line is a good model for the data

Exhibits a pattern, then a line is NOT a good model for the data

Coefficents

R^2 (Coefficient of determination): represents the percentage of the change in the y-variable that can be attributed to its relationship with the x-variable

Ex: r-squared for the regression between x and y is .73, we can say that x accounts for 73% of the variation in y

R (Correlation coefficient): strength of linear line on scale of -1≤0≤1

-1: perfectly linear (negative slope)
0: literally sucks
1: perfectly linear (positive slope)

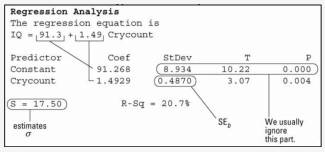
Correlation

X and Y variable assignment doesn't matter

Quantitative values only

Non-resistant (affected by outliers)

Mini Tab!



Scatterplot Vocab

X-variable: explanatory/independent variable (cause)

Y-variable: response/dependent variable (effect)

Extrapolation: predicting outcome outside of the domain

Interpolation: predicting inside the domain [lowest X, highest X]

Sponsored by **Readable.com**
Measure your website readability!
https://readable.com

Sampling Methods

Simple random sample (SRS): every group of objects has equal probability of being selected

Ex: Hat method, calc, table of random digits*

Stratified random sample: sample from each subgroup (good for comparison)

Cluster sample: pick a few subgroups to sample and sample entire subgroup

Systematic Random Sampling: select a sample using a system (like every 3rd)

*ignore number if not in sample, skip if repeat

Bad Sampling

Voluntary response: incomplete data (extremes)

Convenience: chooses easiest individuals to reach

Under-coverage: people aren't reached or accessible

Bad surveying

Non-response: not providing data or talking to you

Response: lying

Poor wording: leans toward bias answer

Principles of Experimental Design

Comparison: Use design that compares two or more treatments

Random Assignment: Use chance to assign experimental units to treatments (balances effects of other variables)

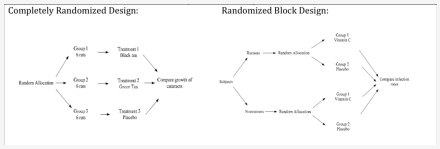
Control: Keep other variables that might affect the response the same for all groups

Principles of Experimental Design (cont)

Replication: Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups

Completely randomized:
-The treatments are assigned to all the experimental units completely by chance
-Control group: that receives an inactive treatment or an existing baseline treatment
-Placebo effect: response to a dummy treatment
-Double-blind experiment: neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received

Experimental Design



Matched pairs:
-Randomized blocked experiment in which each block consists of a matching pair of similar experimental units
-Chance is used to determine which unit in each pair gets each treatment

Law of Large Numbers

As n becomes large the sample mean approaches the population mean

Binomials

1. Each observation falls into one of just two categories –"success" or "failure"

2. The procedure has a fixed number of trials – (n)

3. The observations must be independent – result of one does not affect another

4. The probability of success (p) remains the same for each observation

Geometric

(1-3 same as binomial)

4. The variable of interest is the number of trials required to obtain the first success*

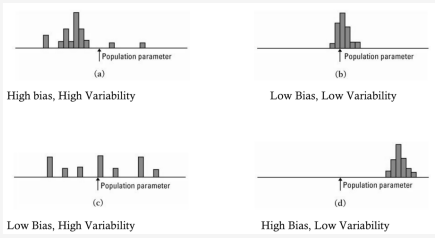
*Geometric is also called a "waiting-time" distribution

Error

Type I: Rejecting the Ho when it is actually true (a false positive); probability = alpha

Type II: Accepting the Ho when it is actually false (a false negative)

Central Limit Theorem



As n becomes very large the sampling distribution for sample mean (x bar) is approximately normal
(n≥30)

Sponsored by **Readable.com**
Measure your website readability!
https://readable.com