Statistics

Definitions

Description	are describing data sets; numbers (*mean*, variance, mode), or pictures (histogram, boxplot)
Inference	*Folgerung,* from the given samples, you make inferences (avg. income of CEO's) or test theories (does an MBA increase income?) about the population
Cases	or individuals (people, cities, stores), things that are measured
Variables	features of the case; income (household), population (city), sales (store)

Random variables RV: New observation or a number, that hasn't yet happened

Continuous	quantitative (numerical) data; salary, weight, age etc.		
	Descriptions:	numerical:	*mean, median*, range, *quantiles*, variance, *SD*
		graphical:	histogram, boxplot
Categorical	nominal (unordered) categories; country of origin, product color ordinal (ordered); small/medium/large		
	Descriptions:	numerical :	frequency tables (how often each value occurs), mode
		graphical :	histogram

In control	A process is in *control* (statistical issue), if it shows no trend in either its *mean* or its *variability*
Capable	A process is *capable* (engineering issue), if its *mean* and *SD* meet the design specifications (follows normal distribution)
Independent	Two *variables* are independent, if knowing the outcome of one gives no additional information about the outcome of the other

Central Limit Theorem: Averages are normally distributed, even if the process isn't

Confidence Interval allows us to quantify just how close we expect the sample average to be to the process mean

Numerical Descriptions:

Measures of location:

Mean	Average value. The sample mean $\xi \cong E(x) \cong$ expected value of a new observation $\cong$ pop. mean μ
Median	Typical Value

Measures of scale:

Quantiles	Measure of spread; *median* is the 50th quantile, quartiles are 25th and 75th quantile	
Variance	Measure of spread s^2; sample numbers 1,3,7,9 -> mean=5 -> deviations from mean -4,-2,2,4 -> squared 16,4,4,16 -> summed = 40 -> divide by n-1; 40/3=13.33 (useful for calculatios)	
	Cross-sectional variation:	Data are a snapshot in time and one variable explains the other; GMAT scores, CEO salaries etc.
	Time series variation:	Trend and seasonality (retail sales etc.); can be eliminated by transformation to relative change: Disadvantage: after transformation, graph does not show trend and increasing variance anymore!
	Random variation:	Lottery, Dices etc.
SD	SQRT(variance s^2)=s; SQRT(13.33) = 3.65 (useful for interpretation) The smaller SD (less variation) the better you can predict a new observation	

Graphical Descriptions:

Descriptive:

Boxplot	one dimension: center, spread, skewness, outliers
Histogram	two dimensions: bar chart of frequencies, center, spread, skewness, bimodality, outliers

Diagnostic:

Transformation If histogram fits normal curve poorly, try to transform data; normality can be tested with the *Normal quantile plot*

Empirical Rule If the normal curve fits well, then: 68% of the data is within +/- 1SD of mean, 95% within 2SD, 99.7% within 3SD

Y-Axis X-Axis	**Quantitative/Continuous**	**Categorical**
Quantitative/Continuous	Correlation and Regression: Scatterplot Example: Do older CEO's make more money than younger CEO's? Issue: How good is the summary equation? Numerical summary: Mathematical equation describing the trend of the scatterplot Graphical summary: Scatterplot	Logistic regression
Categorical	Analysis of variance: Side-by-side Boxplot Example: Do CEO's in some industries make more than others? Issue: How much higher or lower must the group average be before we conclude that CEO's in that group do better or worse than average? Numerical summary: Mean and SD per group Graphical summary: Side by side boxplot	Contingency tables: Crosstabs or Mosaic Plots Example: are MBA's more likely to enter the Financial industry than others? Issue: How different do sample proportions have to be before we conclude that the proportion of MBA's entering Financial industry is different than the proportion entering another industry? Joint relationship: Pr(MBA & Financial) Conditional Relationship: Pr(MBA|Financial) or Pr(Financial|MBA)

Parameters vs. Statistics

	Parameters	Statistics
Mean	μ	ξ
SD	σ	s
Var	σ^2	s^2
Proportion	π	p
Regression slope	β	b

Sampling Distribution: $\mu_{\xi} = \mu$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = SE(\bar{x})$$

Useful Formulas:

E(aX + b)	=	aE(X) + b
E(aX + bY)	=	aE(X) + bE(Y)
Var(aX + b)	=	a^2 Var(X)
Var(aX + bY)	=	if variables dependent: a^2 Var(X) + b^2 Var(Y) + 2ab * Cov(X,Y) if variables independent: a^2 Var(X) + b^2 Var(Y)

Cov(X,Y)	=	Sum of all $((x-\xi)(y-\bar{y}))$	= 0, if X,Y are independent. Cov. cannot be compared the larger, the stronger the covariance
Corr(X,Y)	=	Cov(X.Y)/(SD(X) * SD(Y))	$-1 \leq$ Corr(X,Y) ≤ 1, strength of **linear** relationship comparable, the closer to 1 the stronger the linear relation

E(X*Y)	=	E(X) * E(Y) - Cov(X,Y)
Var	->	$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2$
SD	->	$s = \sqrt{s^2}$

How big sample size = $\frac{1}{(mE)^2}$

Example

X = return on GM
Y = return on IBM
X,Y are independent!

a = b = amount of shares you buy each = 0.5
Total return T = aX + bY

Expected return E(T) = aE(X) + bE(Y)
Var(T) = a^2 Var(X) + b^2 Var(Y)

SD(T) = $\sqrt{\text{a2 Var(X)} + \text{b2 Var(Y)}}$

t-tests

One sample t-test

- One sample
- Two or more samples from same population: for each sample take a one sample t-test and compare

Two sample-test

Basically test, whether the means of two samples lie within the 95% confidence intervals of each other.

- Two independent samples from different populations: are the real means of the two samples the same or different? (H_0: μ_a-μ_b=0)
- Two dependent samples but with unequal sample sizes

Paired (one sample) t-test (Class 8)

- Two dependent samples (two observations taken from the same unit in the sample):
  This is usually cheaper than two independent samples and show less variability!
    - take the difference of each observation
    - calculate SD and ξ of all differences
    - perform a one sample t-test (H_0: mean difference = 0)

Chi-squared test

Tests whether two samples are independent or not (H_0: X and Y independent)

Null hypothesis: existing condition (status quo). Often the option whose choice would lead to no changes.

Alternative hypothesis: The option whose choice would lead to changes, to alter the status quo; switch brands, switch medical treatment, invest in new company etc. Usually the choice you hope to show is true

p-value: is a measure of the credibility of the null-hypothesis (but it is NOT the probability that the null-hypothesis is true, the probability of H_0 cannot be calculated). Small p-values give evidence *against* the null hypothesis
rule of thumb: $p > 0.05$ is considered large

Bias in surveys

Frame coverage bias: Happens when the sampling frame (the frame from which you get your samples) misses important members of the population; women selected from member-lists of women's clubs do not represent all women.

Size bias: The sample is too small or some people are more likely to be included in the survey; people who stay longer in the hotel are more likely to be included in the survey but do not represent the average opinion about the hotel

Non-response bias: Units that do not answer your questions look different than those who do; women who sent back the questionnaire do not represent all the women who have received the questionnaire.

Selection bias: Only units with strong opinions are included

Question sensitivity bias: If the questions are sensitive to social taboos etc., the answers might not be truthful

Reporting bias: Only 'interesting' reports get published; everything gives you cancer

Lurking variables: Does smoking cause cancer or do smoker have a gene which causes both the bad habit and the cancer?

Residuals/Regression

Regression explains variability! How much variability has the regression explained? Am I confident with the relationship between X and Y? Is there a relationship between X and Y? What Y would I predict for a given X?

Residual is the vertical deviation of a point from the fitted regression line.

Variability is **partially** (you also will have to look at the residual plot to check whether there is any trend or pattern) explained in R^2; the bigger the better, and RMSE (SD of residuals); the smaller the better.
$R^2=0$; no linear relationship between X and Y

Regression analysis assumes, that the observations are independent with equal variance. If not:
If Data doesn't follow a straight line, Residual plot shows a bend.
Use transformations and polynomials until pattern in Residual plot is gone.

If Variance is not constant (*heteroscedasticity*), Residual plot shows a *funnel.*
Typically happens, when data are averages (weight proportional to size) or totals (weight inverse to size).
Use transformations (log and sqrt) until pattern in Residual plot is gone.

If Y's are dependent (*autocorrelation*), Residual Plot shows *tracking.*
The past contains information about the future, happens only with time series.
Use transformations (y/4) until pattern in Residual plot is gone.

Scatterplot smoothing (FSW2, p5) attempts to separate the systematic pattern from the random variation.
It allows us to predict new observations!

Extrapolation occurs when you predict new observations **outside** the range of data.
Interpolation occurs when you predict new observations **inside** the range of data (though not necessarily at appoint for which you have data). Interpolation is a lot less riskier than extrapolation.

The *X variable* is the factor that we can manipulate to affect the outcome of Y. X represents what we know and Y is what we want to predict.

Outliers (salary of Walt Disney's CEO);
a point with an unusual Y value (big residual) but **not** an unusual X value;
- little *influence* on slope of regression
- some *influence* on intercept of regression
- big *influence* on residual SD

Leverage (cottage); a point with an unusual X value
leverage is not necessarily bad!
- can have a big *influence* on slope, interception and residual SD of regression

Influence (small on CEO salary and large on cottage);
a point that would change the regression **a lot** if it were removed
If the influence is very big (changes your conclusions), you either can
- make a report about both results
- use transformations and work on a scale where the point is not influential anymore
- delete the point, if:
    - point was recorded in error
    - you only want to use the model to predict 'typical' observations

DO NOT delete points if:
- just because they don't fit the model
- you want to use the model to predict unusual observations (cottages)

Sampling Distributions

Marc Rogivue