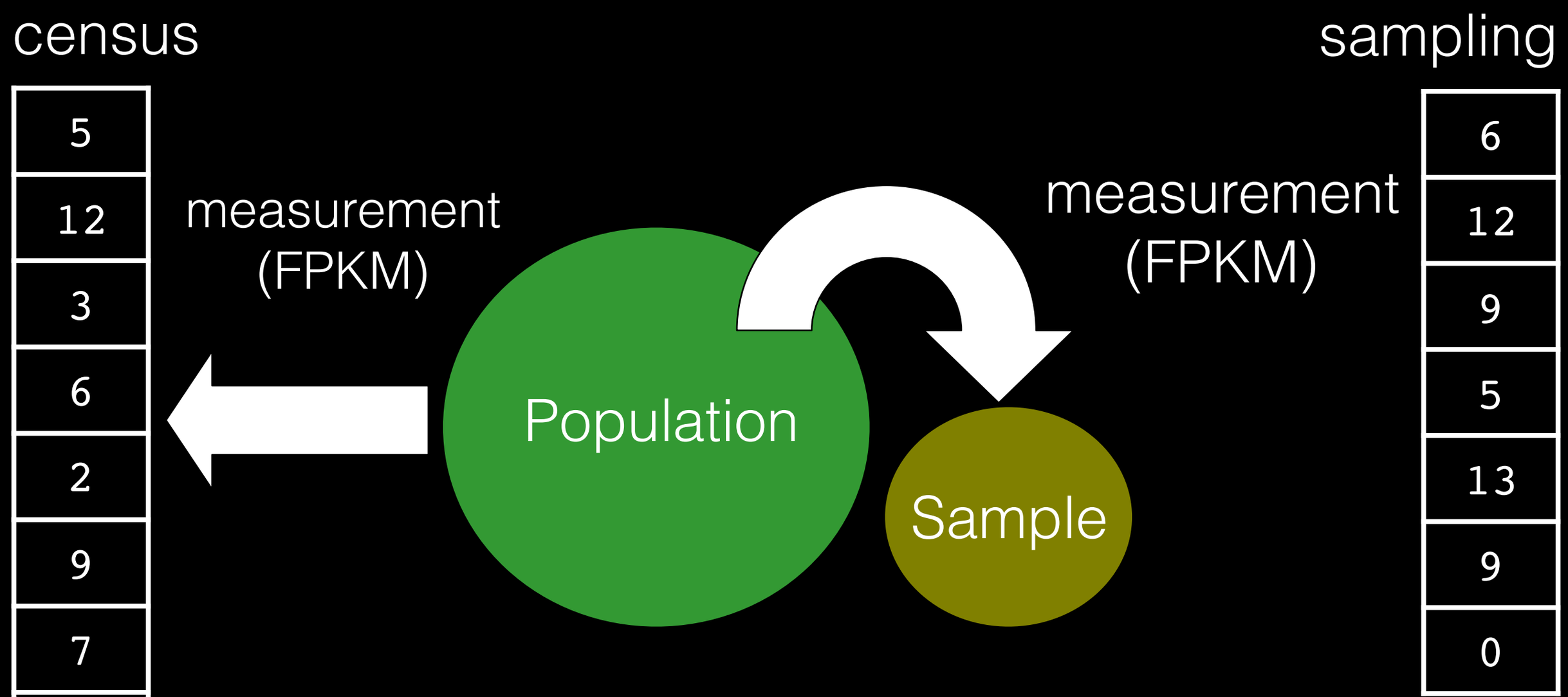
Statistics for Bioinformatics
Introductory Concepts

Katie Pollard

BMI 206

- Sampling
- Estimation
- Data Types
- Association
- Basic Probability
- Hypothesis Testing

Sampling and Study Designs



- Population = All breast cancer patients in the United States
- Sample = 7 random patients from UCSF medical center
- Experimental unit = a patient
- Variable = expression of IL-10 in B cells measured via qPCR

Sampling in Bioinformatics

Definitions not always clear in bioinformatics:

- Sample size might be n=1
- Many variables may be measured
- Variables may be highly correlated

Statistical Objectives

Statistics use data for a few main things

- Estimation: make a best guess about the value (or range of plausible values) of a population parameter
- Testing: make a decision about whether or not a population parameter is some value
- Modeling: quantify relationships between variables (involves estimation, testing) and optionally use model for prediction

Parameter Estimation

Statistics convert data into estimates of population parameters, e.g.

- Univariate: mean, median, variance, skew
- Multivariate: correlation, covariance, regression coefficient, odds ratio, relative risk

What is the error in an estimator?

- Bias
- Variance

Confidence intervals and tests quantify error

Study Designs

Design should reflect the objectives of study

- Observational vs. experimental
- Static vs. longitudinal
- Prospective vs. Retrospective
- Case-control vs. cross-sectional

Study Designs

Some important design considerations:

- Can you generalize to a larger population or a broader context?
- Can you infer causality or only association?
- Was there any bias in collecting the data?
  - Selection bias
  - Nonresponse bias
  - Measurement bias

Data Types

Types of Variables

- Categorical (qualitative)
  - Ordered? Nominal, ordinal, interval
  - Number of levels
- Numerical (quantitative)
  - Discrete (e.g., integer counts)
  - Continuous (e.g., real numbers)
  - Range of values

How is my variable distributed?

Commonly used distributions in bioinformatics:

- Normal/log-normal
- Binomial/product-binomial
- Multinomial/product-multinomial
- Poisson
- Negative Binomial

What data type does each distribution model?

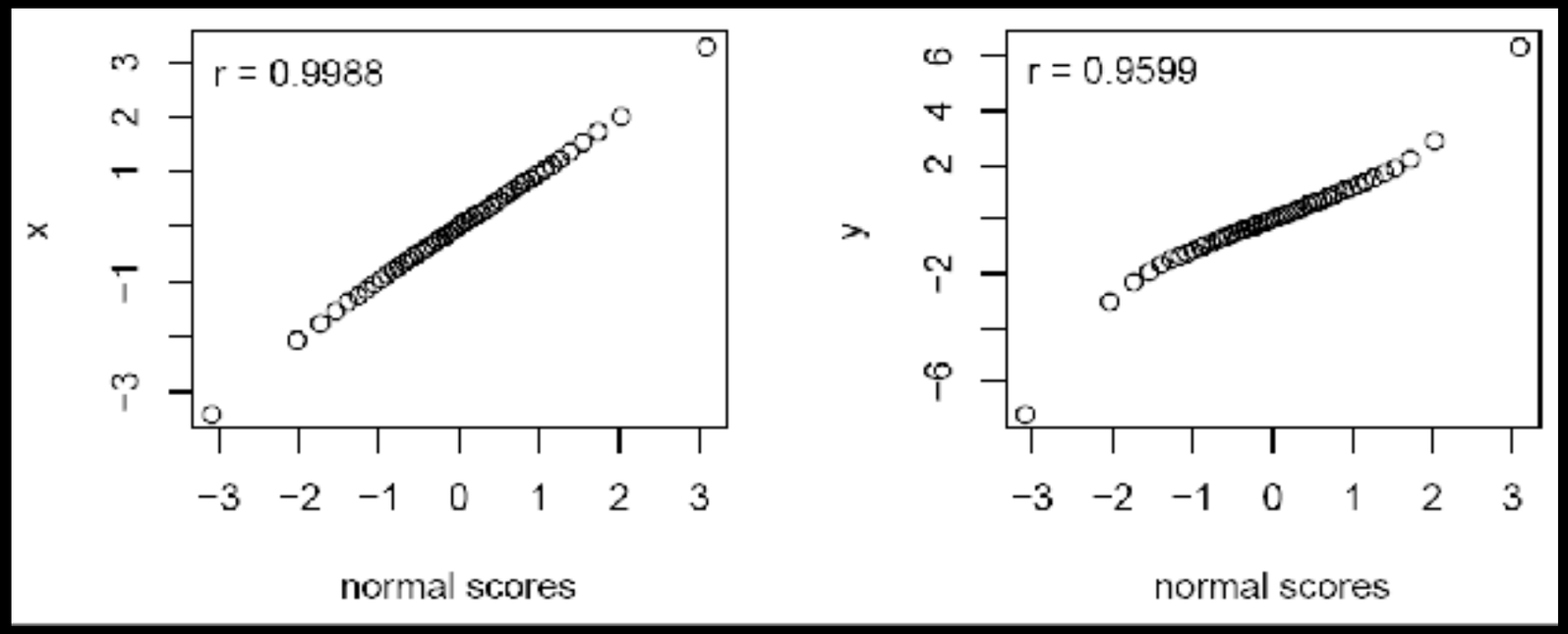
What assumptions do you make using these distributions?

Quantiles

The value *a* such that Pr(X<a)=N% is the Nth percentile, also called the N/100 quantile.

Quantiles of two distributions can be compared to see how different the distributions are.

- Often observed vs. theoretical
- Check for normality or other distribution



Q-Q Plots
Linear if same

Data Transformations

If Y increases a non-constant amount per unit increase in X, transformation may produce a linear relationship:

- Log or exponentiate
- Root or raise to a power
- Reciprocal
- Z-scores (subtract mean, divide by standard error)

For non-continuous data (e.g., counts), other models are typically needed. Generalized linear models will be covered next week.

How Many Variables?

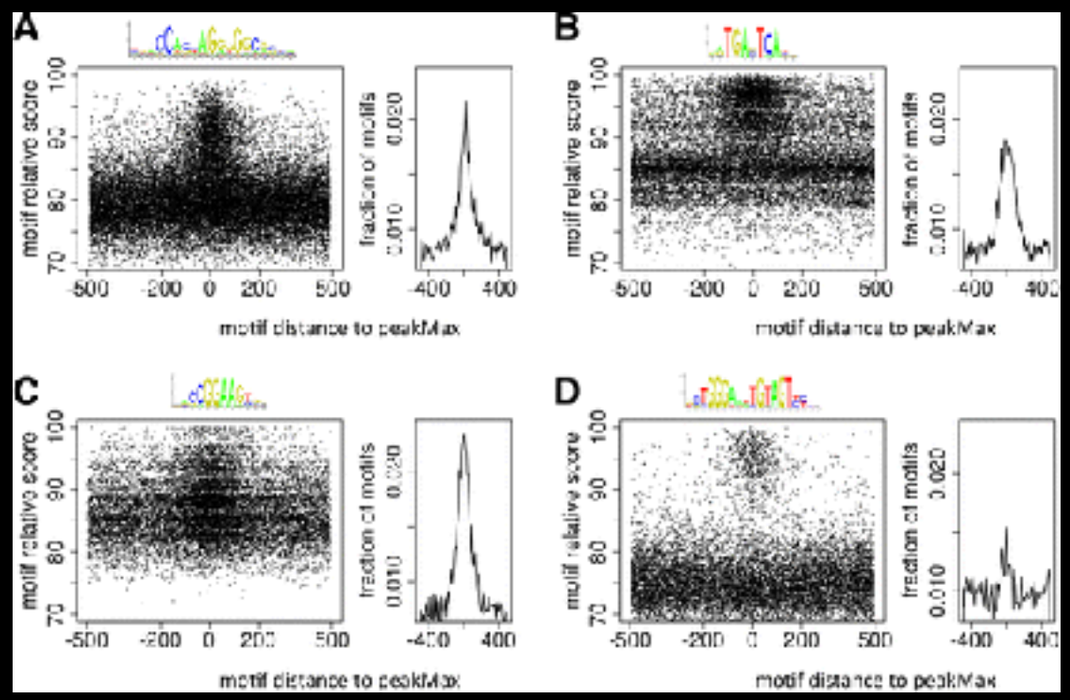
Data is a set of measurements on ≥1 variable.

- Univariate = 1 variable
- Bivariate = Exactly 2 variables
- Multivariate = ≥2 variables

Describes the number of variables measured on each experimental unit.

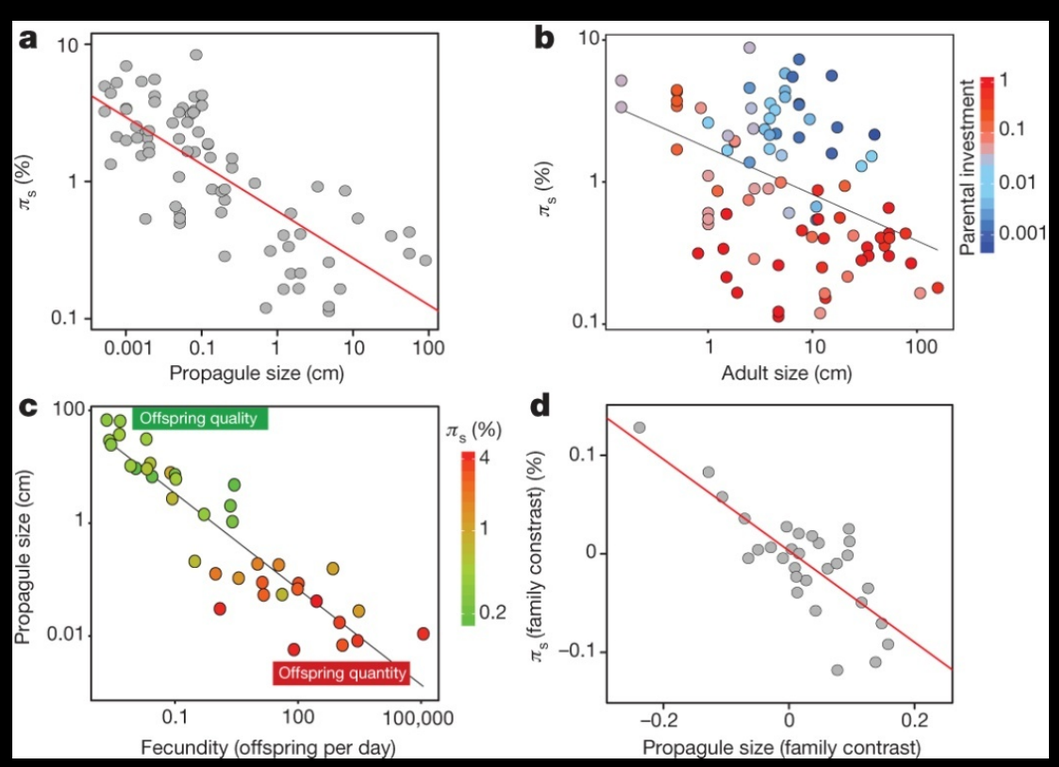
Statistical Association

Statistical Associations and Modeling in Bioinformatics



Zinc finger motifs are enriched in ChIP-seq peaks for non-zinc-finger transcription factors

Hunt & Wasserman (2014) Genome Biology



Life-history traits are correlated with population genetic diversity across animals

Romiguier et al. (2014) Nature

Association

Statistical association is <u>any</u> dependence between two random variables.

- Dependence means that mathematically probabilistic independence is not satisfied.
- Much more general than correlation, e.g.,
  - Measures of association for categorical data
  - Mutual information, dual total correlation, maximal information coefficient
- Neither association nor correlation implies causality.
- Conditional association depends on other variables.

Correlation

Pearson's correlation coefficient (ρ) is estimated by:

$$r = \frac{\sum_{i=1}^{n} z_x(i) z_y(i)}{n-1}$$

- Quantifies <u>linear</u> X vs. Y relationship.
- $-1 \leq r \leq 1$
- Positive ($r>0$) if positive slope
- Negative ($r<0$) if negative slope
- $r=0$ if no <u>linear</u> relationship (may have other relationship)
- Coefficients in linear models measure correlation

Spearman's correlation and Kendall's tau are more robust. They measure rank correlation (monotonicity).

Enrichment

Quantifies excess overlap in sets versus expectation

- Refers to counts of observations in sets
- Not applicable to quantitative data
- Expectation is relative to a null distribution, e.g.,
  - Independence
  - Background level of dependence
- Statistical tests use hypergeometric, binomial, multinomial distributions. Also simulation.

Example: Gene Ontology and RNA-seq

Sets of genes annotated with different ontology terms. For each term, test if genes differentially expressed in cancer vs. healthy are enriched.

Relating Different Data Types

Covariate (dependent variable)

Outcome (independent variable)

	Continuous	Categorical
Continuous	Linear Regression	ANOVA
Categorical	Generalized Linear Model Regression (e.g., Logistic)	Contingency Tables / Log-linear Model Regression

Relationships Between Variables

Variables may play different roles in the study

- Response vs. explanatory (covariate)
- Extraneous vs. variables of interest
- Confounders
- Measured vs. not

Are the variables independent or not?

Probability

Outcomes

Observed data is typically one of many possible "outcomes" or "events"...

- Imagine repeating a random experiment or repeatedly sampling from a population.
- After many repetitions, you would get an idea about which outcomes are most likely to be observed.
- The probability or likelihood of an outcome is its relative frequency in the whole population.
- Likelihood can also be though of as "long-term" frequency after a lot of sampling.

Sample Space

The set of all possible outcomes of a single repetition of a random experiment.

- Sample space is a collection of simple events (outcomes of one repetition of experiment)
- Simple events can involve >1 random variable
- There is a probability associated with every simple event, denoted P(A) for event A
- If events are equally likely,

$$P(A)=1/\{\#\text{ simple events}\}$$

Rules of Probability

- $P(A)$ is a number between 0 and 1.
- $P(A) = 0$ means A never occurs.
- $P(A) = 1$ means A always occurs.
- Probability A does not occur is $P(A^c)=1 - P(A)$
- Sum of the probabilities of all the simple events in the sample space equals 1.

Probability of an Event

Any combination of simple events is an event.

- A simple event is an event.
- The empty set is also an event.
- Probability of an event B is the sum of the probabilities of all the simple events in B.
- If the simple events are equally likely,

  P(B) = {# simple events in B}/{# simple events}
- Counting rules (combinatorics, permutations) help compute these numbers for large or complex sample spaces

Simultaneous Events

Two outcomes are mutually exclusive if they cannot both occur simultaneously

- simple events
- events with no shared simple events

Probability is additive, but must account for simultaneous events if not mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- “or” is the same as “union”
- “and” is the same as “intersection”

DNA Sequence Changes

Example: Probabilities for single DNA base changes

Sample space: {A→A, A→C, A→T, A→G, C→A, C→C, C→T, C→G, T→A, T→C, T→T, T→G, G→A, G→C, G→T, G→G}

- Event {ends A} = {A→A, C→A , T→A , G→A}
- Event {ends T} = {A→T, C→T, T→T, G→T}
- Event {no change} = {A→A,C→C,T→T,G→G}

P(ends A or T) = P(A→A) + P(C→A) + P(T→A) + P(G→A) + P(A→T) + P(C→T) + P(T→T) + P(G→T)

P(no change or ends A) = P(A→A) + P(C→C) + P(T→T) + P(G→G) + P(A→A) + P(C→A) + P(T→A) + P(G→A) - P(A→A)

Conditional Probabilities

Outcomes are independent if their conditional probabilities equals the marginal probabilities:

- Written P(A|B)=P(A). Equivalently, (B|A)=P(B).
- Multiplicative Rule: P(A and B) = P(A|B) P(B)
- Rearranged is Bayes Rule: P(A|B) = P(A and B)/P(B)
- If A and B are independent, P(A and B) = P(A) P(B)
- P(A and B) also written P(A,B) is the joint probability

Probability Estimation

Two methods for computing the probability of an experimental outcome, e.g., $P(X=x)$, $P(X>x)$:

1) Empirically from a large sample (repeat experiment many times same way)
   - Use sample directly to estimate event likelihood
   - Use sample to estimate a parameter and then employ a theoretical distribution to compute complex event probability

2) By simulation (repeat fake experiment many times, must be similar to real situation)

Information Theory

How information is quantified or encoded

- Entropy: uncertainty, average bits needed to store, depends on size of sample space and probabilities of events (CS version of these concepts)

$$H(X) = - \sum_x P(x) \log P(x)$$

- Joint entropy: $H(X,Y) = H(X) + H(Y)$ if X and Y are independent. Else

$$H(X,Y) = - \sum_{x,y} P(x,y) \log P(x,y)$$

- Conditional entropy: $H(X|Y) = H(X,Y) - H(Y)$
- Mutual information: $I(X;Y) = H(X) - H(X|Y)$

Hypothesis Testing

Components of a Hypothesis Test

1. Parameter: quantity of interest
2. Null and alternative hypotheses: statement about parameter value
3. Test statistic: quantify evidence
4. Error rate: control mistakes
5. Null distribution: assess significance
6. Procedure: decision rule

Parameters

Typically, we are interested in testing if a parameter or contrast is zero, e.g.

One group: mean = 0, correlation = 0

Two groups: difference in means = 0

Many groups: all means are equal

Multi-factor: interaction = 0

Tests for categorical data include independence, enrichment, homogeneity

Null Hypothesis

The null hypothesis is a statement of the form

$$H_0\text{: parameter = hypothesized value}$$

- It is a claim about a population characteristic.
- It is the default conclusion, assumed to be true until rejected in favor of an alternative.
- The hypothesized value is typically a single number.

Alternative Hypothesis

The alternative hypothesis is a statement of one of the following forms:

H_a: parameter $\neq$ hypothesized value

Same value as in H_0.

H_a: parameter $>$ hypothesized value

H_a: parameter $<$ hypothesized value

- It is the competing claim, assumed to be false until proven true based on sample data.

Example: Proportion of GC base pairs in DNA

- The following are legitimate hypotheses:

  $H_0: \pi = 0.5$ vs. $H_a: \pi \neq 0.5$

  $H_0: \pi = 0.5$ vs. $H_a: \pi > 0.5$

  $H_0: \pi = 0.5$ vs. $H_a: \pi < 0.5$

- These are not:

  $H_0: \pi = 0.5$ vs. $H_a: \pi = 0.45$

  $H_0: \pi > 0.5$ vs. $H_a: \pi = 0.5$

Statistics

A test statistic is a quantity computed from sample data that is used as the basis for a rejection decision.

- Frequently it is of the form:

(estimate-hypothesized value)/se(estimate)

- How likely would it be to observe this value of the test statistic if H_0 true?

Null Distribution & P-Value

The probability of obtaining a test statistic as large (or larger) than the one observed under a null distribution (i.e., assuming H_0 is <u>true</u>) is called a p-value.

- The p-value is small if the observed statistic would be very unusual under the null.
- The p-value is a single number that summarizes the evidence for/against H_0 in the data.
- If the sample data is inconsistent with H_0, then the test statistic will be large in magnitude (i.e., in the tail of the null distribution) and the p-value will be small.

Example: Proportion of GC base pairs in DNA

- $H_0: \pi = 0.5$ vs. $H_a: \pi \neq 0.5$ "two-sided"

  Reject if the sample proportion p is far from 0.5.

- $H_0: \pi = 0.5$ vs. $H_a: \pi > 0.5$ "greater"

  Reject if p is well above 0.5 (>0.51? >0.75?).

- $H_0: \pi = 0.5$ vs. $H_a: \pi < 0.5$ "less than"

  Reject if p is well below 0.5 (<0.45? <0.3?).

Testing Procedure

- A hypothesis testing procedure is a rule for deciding if you will reject H_0 (or not) based on the observed data (i.e., value of the statistic).
- If the test is conservative, it will tend not to reject H_0 unless the evidence is very strong.
- In this case, you will rarely reject H_0 falsely.
- However, you may often fail to reject H_0 when in fact it is not true (low power).

Testing Procedure

- A rejection decision is of the form:

Reject H_0 if p-value $\leq \alpha$

Fail to reject H_0 if p-value $> \alpha$

- The value α is the significance level of the test, *i.e.* P(Type I error), chosen in advance.

Errors

H_0	Reject	Fail to Reject
True	Type I error	
False		Type II error

- P(Type I error) = α = level of significance
- P(Type II error) = β
- P(reject H_0) = power

If H_0 false, power = 1-P(Type II error) = 1- β

Reject H_0 if p-value $\leq \alpha$.

- If H_0 is true, you have made a Type I error (also known as a "false positive").
- If H_0 is false, you are correct ("true positive")

Fail to reject H_0 if p-value $> \alpha$.

- If H_0 is true, you are correct ("true negative")
- If H_0 is false, you have made a Type II error (also known as a "false negative")

Example: Proportion of GC base pairs in DNA

Test H_0: $\pi = 0.5$ vs. H_a: $\pi > 0.5$

Suppose $\pi = 0.6$ (*i.e.* H_0 is false).

- Rejecting is correct.
- Failing to reject is a Type II error.

Suppose $\pi = 0.5$ (*i.e.* H_0 is true).

- Rejecting is a Type I error.
- Failing to reject is correct.

As significance $\alpha \downarrow$ $\beta \uparrow$, and hence power $\downarrow$

The typical way to deal with the trade-off between Type I and Type II error:

1. Choose the maximum tolerable significance level α based on knowledge of the problem.

2. Then, among all level α tests select the one with the greatest power (*i.e.* lowest β).

The significance level is determined by the cost of making a Type I (vs. Type II) error.

Some methods balance Type I and Type II error.

In addition to the level α of a test, three other factors affect power (for a fixed level α):

- Sample size: as $n \uparrow$ $\beta \downarrow$, so power $\uparrow$.
- Discrepancy between true parameter value and hypothesized value: The farther the true value is from the hypothesized value, the easier it is to detect the difference, so a Type II error is less likely and power $\uparrow$.
- Variance: The more variable the distribution is, the lower power will be for fixed sample size and discrepancy, because the true parameter (and discrepancy) will be estimated with greater error.

Testing Summary

- After collecting sample data, the hypotheses H_0 and H_a are evaluated.
- H_0 is rejected in favor of H_a <u>only</u> if there is sufficient evidence in the sample data to strongly suggest that H_0 is false.
- Else H_0 is <u>not</u> rejected.
- Decision: Reject H_0 vs. fail to reject H_0.

Strong evidence for H_a

No strong evidence against H_0