Applied Statistics

Maureen Hillenmeyer
Stanford University

June 2005

Contents

Abstract

Applied statistics for 2005 quals.

Part I

Descriptive Statistics

Chapter 1

Probability Notation

Notation:

Probability of A =

$$P(A) \ : \ P(A) \geq 0, \ \sum_{A} P(A) = 1$$

Joint probability of A and B =

$$P(A, B)$$

Conditional probability of A given B =

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

Product rule:

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Marginal probability of A given all possible values of B =

$$P(A) = \sum_{B} P(A, B)$$

Independence of A and B:

$$P(A, B) = P(A)P(B)$$

Bayes' Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Combinations:

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{(n-r)!r!}$$

Chapter 2

Statistics of Location

Statistics of location describe the position (e.g. mean). Statistics of dispersion describe the variability (e.g. standard deviation).

In a unimodal, symmetric distribution (e.g. normal), the mean, median, and mode are identical.

2.1 Mean

2.1.1 Arithmetic Mean

The arithmetic mean of X, or $\bar{X}$ is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.1}$$

2.1.2 Geometric Mean

The geometric mean of X is

$$\begin{aligned} GM_X &= \sqrt[n]{\prod_{i=1}^{n} X_i} \\ &= antilog \frac{1}{n} \sum_{i=1}^{n} logX_i \end{aligned} \tag{2.2}$$

When to use the geometric mean? From Wikipedia:

The geometric mean is useful to determine "average factors". For example, if a stock rose 10% in the first year, 20% in the second year and fell 15% in the third year, then we compute the geometric mean of the factors 1.10, 1.20 and 0.85 as (1.10 ? 1.20 ? 0.85)1/3 = 1.0391... and we conclude that the stock rose 3.91 percent per year, on average.

2.1.3 Harmonic Mean

Harmonic mean H_X:

$$\begin{aligned} \frac{1}{H_X} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i} \\ &or \\ H_X &= \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}} \end{aligned} \tag{2.3}$$

When to use the harmonic mean? From Wikipedia:

For instance, if for half the distance of a trip you travel at 40 miles per hour and for the other half of the distance you travel at 60 miles per hour, then your average speed for the trip is given by the harmonic mean of 40 and 60, which is 48; that is, the total amount of time for the trip is the same as if you traveled the entire trip at 48 miles per hour.

2.2 Median

The median of a set of values is the middle value, when they are sorted high to low. If there is an even number of values, the median is the mean of the middle two.

The median is the 50th percentile.

2.3 Percentiles

The percentile is the fraction of points that lie below the given value.

To calculate percentile, first order the values. The 50th percentile, the median, is the value at position $(n+1)/2$.

In general, the pth percentile is

$$\frac{(n+1)}{100/p} \tag{2.4}$$

2.4 Mode

The mode is the value of a set that is most prevalent.

Chapter 3

Statistics of Dispersion

3.1 Range

The range is the difference between the maximum and minimum values of the group.

3.2 Variance

For a population:

Variance σ^2 is

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n} \tag{3.1}$$

For a sample:

Variance s^2 is

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \tag{3.2}$$

3.3 Standard Deviation

Standard deviation is the square root of the variance.

For a population:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}} \tag{3.3}$$

For a sample:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \tag{3.4}$$

3.4 Standard Error

The standard error of the mean is the standard deviation divided by square root of the sample size.

For a population:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{3.5}$$

For a sample:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \tag{3.6}$$

3.5 Interquartile Range

The data points that lie between the 25th and 75th percentile. (I think? double check)

3.6 Coefficient of Variation

The coefficient of variation allows variance comparison between populations with different means. It presents the standard deviation as a percentage of the mean:

$$V = \frac{\sigma \times 100}{\bar{X}} \tag{3.7}$$

3.7 Moment statistics

The rth central moment is the average of deviations of all items from the mean, each raised to the power r:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^r \tag{3.8}$$

The first central moment equals zero:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})$$

The second central moment is the variance:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The third and fourth central moments are used to calculate skewness and kurtosis.

Chapter 4

Boxplots

Boxplots contain

- Median: center line
- Interquartile range (1st and 3rd quartiles): Box
- Extreme values: "Whiskers" (vertical lines). If all values are within 1.5 IQR, then the whiskers only extend to the max/min values.
- Outliers ($>$ 1.5 IQR): Points

Chapter 5

Normalization

5.1 Z-score

5.2 Double-centering

Subtract row and column means, and add back grand mean.

Part II

Distributions

Chapter 6

Discrete

6.1 Binomial

Given a choice of fruits, apple (A) or banana (B), let $P(A) = p$ and $P(B) = q$.

In choosing one fruit, the sample space and corresponding probablities are

$$\{A, B\}$$
$$\{p, q\}$$

In the case of one trial, the variable is a **Bernoulli random variable**.

With two fruits (and AB = BA):

$$\{AA, AB, BB\}$$
$$\{p^2, 2pq, q^2\}$$

And three fruits:

$$\{AAA, AAB, ABB, BBB\}$$
$$\{p^3, 3p^2q, 3pq^2, q^3\}$$

... etc.

The coefficients in the probabilities are equal to the number of ways that the outcome can be obtained.

Binomial expansion summarizes this result:

$$(p + q)^n \tag{6.1}$$

where n is the sample size.

The probability mass function of a binomial distribution is

$$\begin{aligned} P(X = x) &= \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \end{aligned} \tag{6.2}$$

where x is the number of "successes" (here, the number of apples).

The binomial distribution is the expected distribution of outcomes in random samples of size n, with probability p of success.

Mean and variance of binomial distribution:

$$\begin{aligned} \mu &= np \\ \sigma &= \sqrt{npq} \\ \sigma^2 &= npq \end{aligned}$$

6.2 Poisson

The Poisson distribution can be used to approximate the Binomial distribution when one event is rare ($p < 0.1$), and the sample size is large ($np > 5$).

A Poisson variable Y must be

1. **Rare:** Small mean relative to the number of possible events per sample
2. **Random:** Independent of previous occurrences in the sample

This distribution can model the number of times that a rare event occurs, and test whether rare events are independent of each other.

The parameter λ is the expected number of successes. If X is binomial with large n and small p, the number of success is approximately a Poisson random variable with $\lambda = np$.

The probability mass function of a Poisson distribution is

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}. \tag{6.3}$$

λ is the only parameter needed to describe a Poisson distribution. It is equal to both the variance and the mean:

$$\lambda = \mu = \sigma^2. \tag{6.4}$$

Thus the expected frequency of seeing X rare events is

$$e^{-\mu}\frac{\mu^X}{X!}, \tag{6.5}$$

and a simple test of whether a variable is Poisson-distributed is the coefficient of dispersion:

$$CD = \frac{s^2}{\bar{Y}}, \tag{6.6}$$

where s^2 is the sample variance and $\bar{Y}$ is the sample mean. Samples having CD close to one are Poisson-distributed.

The Birthday Paradox

If there are n people in a room, there are $\left(\begin{array}{c} n \\ 2 \end{array} \right)$ pairs of people. We define a success as having one pair share a birthday, with probability 1/365. Thus

$$\begin{aligned} n &= \left(\begin{array}{c} n \\ 2 \end{array} \right) \\ p &= 1/365 \end{aligned}$$

The expected number of successes is

$$\begin{aligned} \mu &= np \\ &= \left(\begin{array}{c} n \\ 2 \end{array} \right)/365 \\ &= n(n-1)/730 \\ &= \lambda \end{aligned}$$

Thus the probability that no two people share a birthday is

$$\begin{aligned} P(X=0) &= e^{-\lambda}\frac{\lambda^0}{0!} \\ &= e^{-\lambda} \\ &= exp\left\{\frac{-n(n-1)}{730}\right\} \end{aligned}$$

If we want to find the number of people for which the probability is less than 0.5:

$$\begin{aligned} exp\left\{\frac{-n(n-1)}{730}\right\} &\leq \frac{1}{2} \\ exp\left\{\frac{n(n-1)}{730}\right\} &\geq 2 \\ n(n-1) &\geq 730\, ln(2), \end{aligned}$$

which is solved with $n \approx 23$, meaning that if there are 23 people in a room, there is a probability of 0.5 that two of them share a birthday.

6.3 Negative binomial

The negative binomial distribution models the number of trials, n, performed until r successes occur.

$$P(X = n) = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} p^r q^{n-r} \tag{6.7}$$

6.4 Geometric

The geometric distribution models the number of trials, n, performed until a success occurs. Note that this is the same as the negative binomial distribution with $r = 1$.

$$P(X = n) = (1-p)^{n-1} p \tag{6.8}$$

The mean and variance of a geometric random variable:

$$\mu = \frac{1}{p} \tag{6.9}$$

$$\sigma^2 = \frac{1-p}{p^2} \tag{6.10}$$

6.5 Hypergeometric

The hypergeometric distribution is equivalent to the binomial, in that it models the number of successes in n trials, but it accounts for sampling *without* replacement.

Imagine an urn containing N balls: m are white and $N - m$ are black. The hypergeometric distribution is given by

$$P(X = i) = \frac{\begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} N-m \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{6.11}$$

where
m = possible successes = number of white balls in the urn
i = successes = number of white balls selected
n = sample size = total balls selected
N = population size = total balls in urn

If the number of white and black balls (successes and failures) are not explicitly known, but the probabilities are, then:

$$P(X = i) = \frac{\binom{pN}{i}\binom{qN}{n-i}}{\binom{N}{n}} \tag{6.12}$$

where
p = probability of success
q = probability of failure
i = number of successes
n = sample size
N = population size

Thus parameters needed to describe a hypergeometric distribution are
N, the population size,
n, the sample size, and
m, the number of successes (or p, the probability of success).

6.6 Zeta / Zipf

The probability mass function of the Zipf distribution is

$$P(X = k) = \frac{C}{k^{\alpha+1}} \tag{6.13}$$

Chapter 7

Continuous

7.1 Uniform

The probability density function for a uniform random variable X on interval (α, β) is

$$f(X) = \begin{cases} \dfrac{1}{\beta - \alpha} & if \ \alpha < x < \beta \\ 0 & otherwise \end{cases} \tag{7.1}$$

The mean and variance of X:

$$\begin{aligned} \mu &= \frac{\beta + \alpha}{2} \\ \sigma^2 &= \frac{(\beta = \alpha)^2}{12} \end{aligned}$$

7.2 Normal

Parameters: μ, σ^2

The normal probability density function is

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} \tag{7.2}$$

The curve is symmetric about the mean.

$$\begin{aligned} \mu \pm \sigma &\text{ contains } 68.3\% \text{ of the items} \\ \mu \pm 2\sigma &\text{ contains } 95.5\% \text{ of the items} \\ \mu \pm 3\sigma &\text{ contains } 99.7\% \text{ of the items} \end{aligned}$$

and

50% of the items lie within $\mu \pm .67\sigma$
95% of the items lie within $\mu \pm 1.96\sigma$
99% of the items lie within $\mu \pm 2.58\sigma$

We can fit a normal distribution to an observed frequency distribution:

$$Z = \frac{ni}{s\sqrt{2\pi}} e^{-(Y-\bar{y})^2/2s^2}, \tag{7.3}$$

where n is the sample size and i is the class interval of the frequency distribution.

We can also calculate expected frequencies for a normal distribution having the observed mean and standard deviation of the observed sample.

Some distributions (e.g. binomial/multinomial) can be approximated by a normal distribution as the sample size becomes large.

7.2.1 Skewness and Kurtosis

Skewness is the amount of asymmetry in a distribution. In a skewed distribution, the mean and median are not identical.

Skewness of the population is γ_1, and skewness of the sample is g_1.

$$g_1 = \frac{1}{ns^3} \sum_{i=1}^{n} (Y_i - \bar{Y})^3 \tag{7.4}$$

which is the third central moment divided by the cubed standard deviation.

Kurtosis describes the peakedness of a distribution. "Leptokurtic" curves have long tails and "platykurtic" curves have short tails. "Mesokurtic" distributions have the same kurtosis as the normal distribution.

Kurtosis of the population is γ_2 and skewness of the sample is g_2.

$$g_2 = \frac{1}{ns^4} \sum_{i=1}^{n} (Y_i - \bar{Y})^4 - 3 \tag{7.5}$$

which is the fourth central moment divided by the fourth power standard deviation, minus 3.

In the normal distribution, γ_1 and γ_2 are zero. Negative g_1 indicates left skewness, and positive g_1 indicates right skewness. Negative g_2 indicates platykurtosis, and positive g_2 indicates leptokurtosis.

7.2.2 Central Limit Theorem

The Central Limit Theorem states that as sample size increases, the means of samples drawn from a population having any distribution will approach the normal distribution.

As the number of samples increases, the Central Limit Theorem states that the:

- distribution of sample means will approximate normal, regardless of the original distribution
- mean value of sample means will equal the population mean
- standard deviation of the sample means (standard error of the mean) depends on population standard deviation and sample size.

There are many central limit theorems at various levels of abstraction. (E.g. one in Rice book assumes the existence of moment-generating functions, which only exist if the expectation converges.) Central limit theorems are still an active area of research in probability theory.

7.3 t-distribution

This distribution was described by W.S. Gossett, under the pseudonym "Student", so it is sometimes called the Student's distribution.

The deviates $\bar{Y} - \mu$ of sample means from the true means of a normal distribution are also normally distributed. These deviates divided by the true standard deviation, $(\bar{Y} - \mu)/\sigma$ are still normally distributed, with $\mu = 0$ and $\sigma = 1$ (standard normal).

But the distribution of deviates of i samples, each with mean Y_i and standard error $s_{\bar{Y}i}$,

$$\frac{(\bar{Y}_i - \mu)}{s_{\bar{Y}i}} \tag{7.6}$$

is not normally distributed. It is wider because the denominator is the sample standard error instead of the population standard error. It will sometimes be smaller, sometimes larger than expected, so the variance is greater.

The expected distribution of this ratio follows the t-distribution.

The t-distribution's shape is dependent on the degrees of freedom, $n - 1$, where n is the sample size. As the degrees of freedom increase, the t-distribution approaches the normal distribution, and is equal to it when $n = \infty$ (and close to it when $n \approx 25$).

7.4 Gamma

Parameters: n, λ

The gamma distribution models the time until a total of n events has occurred.

Note that this is the continuous equivalent of the negative binomial distribution.

The gamma distribution can model time-to-first-failure events. It is often used for a system with "standby" backups, each having exponential lifetimes with parameter λ.

The probability distribution is

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{(n-1)}}{\Gamma(n)} \quad \text{for } x \geq 0 \tag{7.7}$$

where

$$\begin{aligned} \Gamma(t) &= \int_0^\infty e^{-y} y^{n-1} dy & (7.8) \\ &= (n-1)! & (7.9) \end{aligned}$$

for integral values of n.

7.5 Exponential

Parameters: λ

The exponential distribution models the amount of time until an event occurs.

Note that this is the continuous equivalent of the geometric distribution.

The probability density function is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \tag{7.10}$$

Note that the exponential distribution is the same as the gamma distribution with parameters $(1, \lambda)$. (I.e. time until first event, $n = 1$.)

The mean and variance of x:

$$\begin{aligned} \mu &= \frac{1}{\lambda} \\ \sigma^2 &= \frac{1}{\lambda^2} \end{aligned}$$

The cumulative distribution function is

$$F(x) = 1 - e^{-\lambda x} \tag{7.11}$$

The exponential distribution is **memoryless**:

$$P(X > t + s \mid X > t) = P(X > s) \text{ for all } s, t \geq 0, \tag{7.12}$$

meaning that if the instrument is alive at time t, the probability of survival to time $t + s$ (i.e., from time t to s) is the same as the initial probability of surviving to time s. (I.e., the instrument doesn't remember that it already survived to t).

7.5.1 Laplace distribution

(a.k.a. double exponential)

A variation on the exponential distribution. It arises when a random variable is equally likely to be positive or negative, and it has an absolute value that is exponentially distributed.

7.5.2 Hazard rate function

(a.k.a. failure rate)

The hazard rate function $\lambda(t)$ is

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)}, \tag{7.13}$$

where $\overline{F}(t) = 1 - F$.

$\lambda(t)$ represents the conditional probability that an item will fail, given that it survived until time t. If the lifetime distribution is exponential, then by the memoryless property (the lifetime doesn't affect the probability of failure), the probability of failure should be constant:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{\overline{F}(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-}\lambda t} \\ &= \lambda \end{aligned}$$

7.6 Chi-Square

Parameters: n

The chi-square distribution models the square of the error in n-dimensional space, assuming the coordinate errors are independent standard normal.

If Y_i are standard (mean 0, variance 1) normal independent variables, then

$$\sum_{i=1}^{n} Y_i^2 \tag{7.14}$$

follows a χ^2 distribution with n degrees of freedom.

The χ^2 distribution is the same as the gamma function with parameters $(\frac{n}{2}, \frac{1}{2})$.

7.7 Weibull

The Weibull distribution is often used as the distribution of the lifetime of an item under the "weakest link" model. E.g., if an item fails when one of its many parts fails.

7.8 Beta

The beta distribution models variables whose values fall in an interval $[c, d]$, which can be transformed into the interval $[0, 1]$.

The beta distribution is related to the gamma distribution.

7.9 Dirichlet

Multivariate generalization of the beta distribution.

7.10 F-distribution

Parameters: n, m

The F-distribution models whether two distributions have the same variance.

The F-statistic is defined as

$$F_{n,m} \equiv \frac{\chi_n^2/n}{\chi_m^2/m} \tag{7.15}$$

where χ_m^2 and χ_n^2 are independent chi-squared variables, with m and n degrees of freedom.

Part III

Hypothesis Testing

Type of experiment					
Scale of measurement	Two treatment groups (diff indiv)	Three or more treatment (diff indiv)	Before and after a single treatment (same indiv)	Multiple treatments (same indiv)	Association between two variables
Interval (norm dist)	Unpaired t-test	ANOVA	Paired t-test	Repeated-measures ANOVA	Linear regression or Pearson corr
Nominal	Chi-square	Chi-square	McNemar's	Cochrane Q	Relative rank or odds ratio
Ordinal	Mann-Whitney	Kruskal-Wallis	Wilcoxon signed-rank	Friedman statistic	Spearman rank corr
Survival time	Log-rank or Gehan's test				

Table 7.1: Summary of some statistical methods to test hypotheses (From Glantz 2001)

Chapter 8

Errors and significance

8.1 Null hypothesis

The null hypothesis (H_0) is assumed true until shown otherwise. It typically presumes that there is no difference in the data under examination (e.g. there is no difference in the means of two populations). If a statistical test finds a significant difference, then we reject the null hypothesis (e.g. declare a difference between the means).

8.2 Type I / Type II errors

	True state of null hypothesis	
Statistical decision	H_0 true	H_0 false
Reject H_0	Type I error α	Correct $1-\beta$
Don't reject H_0	Correct $1-\alpha$	Type II error β

Table 8.1: Summary of errors

Type I error occurs if the null hypothesis is rejected when it is true (false positive).

Type II error occurs when the null hypothesis is false but not rejected (false negative).

$$\text{P(Type I error)} = \alpha \tag{8.1}$$

$$\text{P(Type II error)} = \beta \tag{8.2}$$

Graphical illustration of Type I vs Type II errors:
http://www.psychstat.smsu.edu/introbook/sbk26m.htm

8.2.1 The meaning of p-value

The p-value is the probability of type I error: the probability of being wrong when concluding that a difference exists.

8.2.2 Power

α is the probability of a Type I error: that you will wrongly reject the null (when it is true). β is the probability of a Type II error: that you will wrongly accept the null (when it is false).

Power is $1 - \beta$.

In intuitive terms: Assume there is a difference; will you be able to detect it? Power is the chance of detecting a true difference (getting a significant p-value) assuming the given parameters.

Power depends on

1. Sample size. Larger sample size means greater power.
2. Size of difference worth detecting, with respect to the variance. (This must be specified in advance.) is harder to detect smaller differences.
3. α : more stringent cutoff means reduced power.

Relationship between β and α: [Figure of two curves: one under null, and the true distribution, given that there is a true effect. To the right of α on the null curve is the chance of Type I error; to the right of α on the true curve is the power, $1 - \beta$, the chance that we will detect the true difference. To the left is β, the chance that we will make a Type II error.]

Relationship between mean differences and standard deviation:
Define a **noncentrality parameter**:

$$\phi = \frac{\delta}{\sigma} \tag{8.3}$$

where
$\delta = \mu_1 - \mu_2$
σ = population standard deviation.

An increase in σ decreases the power, and an increase in δ increases the power.

If a study doesn't find a significant difference, that does not mean that one does not exist. It may not have had enough power to detect a true difference.

8.3 Confidence intervals

The standard deviate of a sample mean from the true mean is

$$\frac{(\overline{Y}-\mu)}{\sigma_{\bar{Y}}}, \tag{8.4}$$

where $\sigma_{\bar{Y}}$ is $\sigma/\sqrt{n}$.

If the standard deviates are normally distributed, 95% of them will lie between -1.96 and 1.96. (See discussion of normal distribution in section 7.2.)

$$\begin{aligned} P\left(-1.96 \leq \frac{(\overline{Y}-\mu)}{\sigma_{\bar{Y}}} \leq 1.96\right) &= 0.95 \\ P\left(\overline{Y}-1.96\sigma_{\bar{Y}} \leq \mu \leq \overline{Y}+1.96\sigma_{\bar{Y}}\right) &= 0.95 \end{aligned} \tag{8.5}$$

(8.6)

We define these limits as confidence limits,

$$\begin{aligned} L_1 &= \overline{Y}-1.96\sigma_{\bar{Y}} \\ L_2 &= \overline{Y}+1.96\sigma_{\bar{Y}} \end{aligned}$$

These confidence limits L_1 and L_2 can be calculated from the normal distribution if the true parametric standard deviation is known (or if the sample size is large). But if it is not known, it must be calculated from the sample, and thus the confidence limits should come from the t-dstribution.

For confidence limits of probability $1-\alpha$,

$$\begin{aligned} L_1 &= \bar{Y}-t_{\alpha[n-1]}s_{\bar{Y}} \\ L_2 &= \bar{Y}+t_{\alpha[n-1]}s_{\bar{Y}} \end{aligned}$$

where $t_{\alpha[n-1]}$ is the value of the t-distribution at level α (e.g. 0.05 for 95% confidence) with $df = n-1$.

Thus,

$$\begin{aligned} P(L_1 \leq \mu \leq L_2) &= 1-\alpha \\ P(\bar{Y}-t_{\alpha[n-1]}s_{\bar{Y}} \leq \mu \leq \bar{Y}+t_{\alpha[n-1]}s_{\bar{Y}}) &= 1-\alpha \end{aligned} \tag{8.7}$$

For example, at 95% confidence ($\alpha = .05$) and $n = 10$, ($df = 9$), we have

$$\begin{aligned} L_1 &= \bar{Y} - 2.3 s_{\bar{Y}} \\ L_2 &= \bar{Y} + 2.3 s_{\bar{Y}} \end{aligned}$$

To reduce the width of the confidence interval, the standard error of the mean must be reduced. This can be achieved by reducing σ or increasing n, the sample size. (Larger degrees of freedom lead to smaller values of t.)

8.4 One-tailed vs. two-tailed tests

It is easier to reject the null with a one-tailed test than two-tailed test.

A one-tailed test is used when we predict the direction of the difference in advance (e.g. one mean will be larger than the other). With that assumption, the probability of incorrectly rejecting the null is only calculated from one tail of the distribution. In standard testing, the probability is calculated from both tails. Thus, the p-value from a two-tailed test (p_2) is twice the p-value of a one-tailed test (p_1).

$$p_2 = 2\, p_1$$

It is rarely correct to perform a one-tailed test; usually we want to test whether any difference exists.

8.5 Parametric vs. Non-parametric

Use nonparametric methods if the data are not normally distributed or do not meet other assumptions of a given test (e.g. equal variance in all groups).

Chapter 9

Tests for distributions

9.1 QQ / Probability Plots

A Q-Q (quantile-quantile) plot can illustrate whether two groups have a common distribution.

A quantile is the fraction of points below a threshold. At the .5 quantile, half of the data points fall below the threshold and half fall above.

If two groups have a common distribution, plotting their quantiles against each other should form an approximately straight line at a 45-degree angle.

Probability plots compare a data set to a known distribution. They are often used to test normality assumptions.

The normal probability plot is covered here:
http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

9.2 Anderson-Darling

9.3 Shapiro-Wilk

Test for normality.

9.4 KL Divergence

(From Amit)

How similar are two probability distributions?

$$D(p||q) = \sum p(x) \times log((p(x)/q(x)))$$

Note that this is not symmetric.

9.5 Bootstrap to estimate distributions

(From http://en.wikipedia.org/wiki/Bootstrap)
Invented by Brad Efron, further developed by Efron and Tibshirani.

It is a method for estimating the distribution by sampling with replacement. The original sample is duplicated many times, representing a population. Samples are drawn from this population, with replacement.

Sampling with replacement is more accurate than sampling without replacement.

Chapter 10

Differences between two groups

10.1 T-test

For continuous variables.

The t-test examines whether the means of two groups are different.

The t-statistic is

$$\begin{aligned} t &= \frac{\text{difference in sample means}}{\text{variability of sample means}} \\ &= \frac{\overline{X} - \overline{Y}}{SE(\overline{X} - \overline{Y})} \end{aligned} \tag{10.1}$$

where $SE(\overline{X} - \overline{Y})$ can be calculated from individual variances,

$$SE(\overline{X} - \overline{Y}) = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \tag{10.2}$$

or pooled variances (here assuming same sample size n),

$$SE(\overline{X} - \overline{Y}) = \sqrt{\frac{s^2}{n} + \frac{s^2}{n}} \tag{10.3}$$

where s^2 is averaged from the two samples:

$$s^2 = \frac{s_X^2 + s_Y^2}{2}$$

Note that pooling variance can increase sensitivity of a test.

The degrees of freedom for a t-test:

$$df = n_1 + n_2 - 2 \tag{10.4}$$

From the t-statistic, we can compute a p-value using the t-distribution. This p-value is the probability that we reject the null when it is true (probability that we are wrong; type I error).

Note: It can be shown that the t-test is an ANOVA, with

$$F = t^2 \tag{10.5}$$

where F is the F-statistic from section 11.1.

The t-test should not be used to directly compare more than two groups. FIrst ANOVA should be used to see whether *any* difference exists; if one does, then t-tests can be performed to find it. However, multiple hypothesis correction should be applied (see section 16).

10.2 Mann-Whitney U (Wilcoxon) rank sum test

For ordinal (rank) variables. Nonparametric method.

Steps:

- Combine the two samples X_1 and X_2, and order the values by rank.
- Sum the ranks in each group, R_1 and R_2 (actually only one is necessary; see below)
- Compute test statistic U

$$U = n_1 n_2 + \frac{n_1(n_2+1)}{2} - R_1 \tag{10.6}$$

At the extreme, R_1 is equal to $n_1(n_2+1)/2$, and the maximum U will be $n_1 n_2$.

Compare the U-statistic to a U-distribution, which can be approximated by the normal distribution with sample sizes greater than about 20.

10.3 Nominal / categorical variables

10.3.1 z-statistic

Parametric method of analyzing independent Bernoulli trials. If we examine whether there difference in sample proportions (fractions).

	Disease	No disease	Row totals
Treatment	O_{11}	O_{12}	R_1
Control	O_{21}	O_{22}	R_2
Column totals	C_1	C_2	N

Table 10.1: Example 2 x 2 contingency table

$$\begin{aligned} z &= \frac{\text{difference in sample proportions}}{\text{variability of sample proportions}} \\ &= \frac{\hat{p_1} - \hat{p_2}}{SE(\hat{p_1} - \hat{p_2})} \end{aligned} \tag{10.7}$$

where
(**********finish)

Assumptions (which make the method parametric):

- Each trial has two mutually exclusive outcomes
- Probability p of given outcome remains constant
- All trials are independent

(From Glantz 2001)

10.3.2 Chi-square

Nonparametric method. (Assumes nothing about parameters of population.)

Compare observed contingency table to expected contingency table.

Test statistic χ^2 is defined as

$$\chi^2 = \sum_{cells} \frac{(O - E)^2}{E}$$

where O = observed individuals in a cell (see Table 10.2)
and E = expected number of individuals in a cell (see Table 10.3)

The expected value of a cell (i,j), $E(O_{i,j})$, is the probability of each column multiplied by the row total. The probability of each column is simply the column total over the grand total.

	Disease	No disease	Row totals
Treatment	O_{11}	O_{12}	R_1
Control	O_{21}	O_{22}	R_2
Column totals	C_1	C_2	N
Column probabilities	C_1/N	C_2/N	

Table 10.2: Observed contingency table

$$\begin{aligned} E(O_{ij}) &= P(column\ i) \times R_i \\ &= \frac{C_i}{N} \times R_i \end{aligned} \tag{10.8}$$

(Or, equivalently, the probability of each row multiplied by the column total.)

	Disease	No disease	Row totals
Treatment	$(C_1/N) \times R_1$	$(C_2/N) \times R_1$	R_1
Control	$(C_1/N) \times R_2$	$(C_2/N) \times R_2$	R_2
Column totals	C_1	C_2	N

Table 10.3: Expected contingency table

The degrees of freedom are

$$df = (r-1)(c-1) \tag{10.9}$$

where r = number of rows, and c = number of columns.

10.3.3 Fisher's exact test

Alternative to chi-square test. It computes exact probabilities based on the hypergeometric distribution.

It should be used when the sample size is small.

The probability of observing a given contingency table can be computed as

$$P = \frac{\frac{R_1!R_2!C_1!C_2!}{N!}}{O_{11}!O_{11}!O_{11}!O_{11}!} \tag{10.10}$$

To calculate the probability of a table that *at least* that extreme, compute the probabilities for all tables with the same row/column totals, but that have more extreme values. Sum these probabilities.

10.3.4 Relative Risk / Odds Ratios

Odds is defined as

$$odds = \frac{p}{1-p} \tag{10.11}$$

E.g. if 9/10 people in a room are male, the odds of being male is .9 / .1 = 9, or 9 to 1. The odds of being female are .1/.9 = .11, or 1 to 9.

Note:
In discussing relative risk and odds ratio, be careful with the term "control". For relative risk, "control" generally means no treatment, and for odds ratio, "control" generally means no disease.

	Disease (cases)	No disease (control)	Row totals
Exposure (treatment)	a	b	a + b
No-exposure (no-treatment)	c	d	c + d
Column totals	a + c	b + d	

Table 10.4: RR / OR table

The relative risk quantifies the relationship between exposure and outcome:

$$RR = \frac{\text{probability of disease in exposure group}}{\text{probability of disease in no-exposure group}} \tag{10.12}$$

Referring to table 10.4, this becomes

$$\begin{aligned} RR &= \frac{\text{exposed and diseased / all exposed}}{\text{no-exposure and diseased / all no-exposure}} \\ &= \frac{a/a+b}{c/c+d} \end{aligned} \tag{10.13}$$

Analogously, the odds ratio compares the risk of exposure in cases (disease) to controls (no-disease):

$$OR = \frac{\text{odds of exposure in cases (disease)}}{\text{odds of exposure in controls (no-disease)}}$$

The odds of exposure in cases (disease) is:

$$\begin{aligned} &= \frac{\text{treatment and diseased / all diseased}}{\text{no-treatment and diseased / all diseased}} \\ &= \frac{\text{treatment and diseased}}{\text{no-treatment and diseased}} \\ &= \frac{a}{c} \end{aligned} \tag{10.14}$$

The odds of exposure in controls (no-disease) is:

$$\begin{aligned} &= \frac{\text{treatment and no-disease / all no-disease}}{\text{no-treatment and no-disease / all no-disease}} \\ &= \frac{\text{treatment and no-disease}}{\text{no-treatment and no-disease}} \\ &= \frac{b}{d} \end{aligned} \tag{10.15}$$

So the OR is:

$$\begin{aligned} OR &= \frac{a/c}{b/d} \\ &= \frac{ad}{bc} \end{aligned}$$

10.4 QQ / Probability Plots

Covered in section 9.1.

Chapter 11

Differences between three or more groups

11.1 ANOVA

Intuitive ANOVA:

When examining multiple groups, the F-statistic compares the variance between the groups (s_{bet}) to the variance within the groups (s_{wit}).

$$F = \frac{s_{bet}}{s_{wit}} \tag{11.1}$$

If the groups come from the same population (as stated by the null hypothesis), these two variances should be approximately equal, and their ratio should be near one. If the groups are very different, coming from different populations, then the between-group variance will be larger than the within-group variance, and the ratio will be greater than one.

At large F, we reject the null hypothesis and conclude that a difference exists between groups.

One-way (single factor) ANOVA assumes one underlying factor causing the difference.

There are two degrees-of-freedom parameters in ANOVA: ν_n and ν_d. ν_n is the between-groups dof (numerator), and ν_d is the within-groups dof (denominator).

$$\begin{aligned} \nu_n &= m - 1 \\ \nu_d &= m(n-1) \end{aligned}$$

Mathematical explanation of ANOVA:

Organize data into a table of subjects, s_1 to s_n, by treatments, t_1 to t_m.

Within-group variance

Let SS_t be the sum of squares for treatment t, over all subjects s who received t:

$$SS_t = \sum_s (X_{ts} - \overline{X}_t)^2 \tag{11.2}$$

The variance within one treatment group t is thus

$$s_t^2 = \frac{SS_t}{n-1}$$

The within-group sum of squares over all treatment groups t is

$$SS_{wit} = \sum_t SS_t$$

And the estimated within-group variance of the population is the average within-group variance of each treatment:

$$\begin{aligned} s_{wit}^2 &= \frac{s_t^2}{m} \\ &= \frac{\sum_t SS_t}{m(n-1)} \\ &= \frac{SS_{wit}}{DF_{wit}} \\ &= MS_{wit} \end{aligned}$$

where MS_{wit} is the within-groups **mean-square**.

Between-group variance

Let SS_{bet} be the between-groups sum of squares.

$$SS_{bet} = n \sum_t (\overline{X}_t - \overline{X})^2$$

where
$\overline{X}$ = mean over all samples
$\overline{X}_t$ = mean within treatment group t
n = number of samples per treatment.

So

$$\begin{aligned} s^2_{bet} &= \frac{SS_{bet}}{m-1} \\ &= \frac{SS_{bet}}{DF_{bet}} \\ &= MS_{bet} \end{aligned}$$

F-statistic

The F-statistic is computed as

$$F = \frac{MS_{bet}}{MS_{wit}} \tag{11.3}$$

Total variance

The total sum of squares is

$$\begin{aligned} SS_{tot} &= SS_{bet} + SS_{wit} \\ &= \sum_t \sum_s (\overline{X}_{ts} - \overline{X})^2 \end{aligned} \tag{11.4}$$

The total degrees of freedom is

$$\begin{aligned} DF_{tot} &= DF_{bet} + DF_{wit} \\ &= (m-1) + m(n-1) \\ &= mn - 1 \end{aligned} \tag{11.5}$$

And the total variance is

$$s^2 = \frac{SS_{tot}}{DF_{tot}} \tag{11.6}$$

11.2 Kruskal-Wallis statistic

Ordinal (rank) variables. Nonparametric method.

A generalization of the Mann-Whitney rank sum test.

Steps:

- Rank all variables regardless of group
- Compute the rank sum for each group
- Compute test statistic H

$$H = \frac{12}{N(N+1)} \sum_t n_t (\overline{R}_t - \overline{R})^2 \tag{11.7}$$

Chapter 12

Before and after treatment per subject

12.1 Paired t-test

For continuous variables.

Measurements before and after a single treatment.

The t-statistic is

$$t = \frac{\overline{d} - \delta}{s_{\overline{d}}} \tag{12.1}$$

where
δ is the true mean change before and after treatment
$\overline{d}$ is the estimated mean change before and after treatment
$s_{\overline{d}}$ is the standard error of the mean

To test the null hypothesis that $\delta = 0$,

$$t = \frac{\overline{d}}{s_{\overline{d}}} \tag{12.2}$$

12.2 Wilcoxon signed rank test

Ordinal (rank) variables.

Nonparametric version of paired t-test. Measurements before and after a single treatment.

Steps:

- Compute before/after treatment differences in each subject
- Rank differences according to magnitude (ignoring sign)
- Attach sign of difference to each rank
- W = sum of signed ranks

Compare W to known distribution to get p-value.

12.3 McNemar's test for changes

For nominal (binary) variables.

Analogous to paired t-test (two treatments per subject, or before/after).

	Treatment 2	
Treatment 1	+	−
+	a	b
−	c	d

Table 12.1: McNemar table

We cannot use chi-square because it tests the hypothesis that the rows and columns are independent. Here, they are not independent because the same individual is being represented.

The subjects who respond positive to both treatments (cell a) and those who respond negative to both treatments (cell d) are not informative. We are interested in the subjects who respond differently to the two treatments (cells b and c).

We compute the chi-square statistic for these two cells

$$\chi^2 = \sum_{\text{cells b,c}} \frac{(O-E)^2}{E} \tag{12.3}$$

with one degree of freedom.

Chapter 13

Multiple treatments per subject

13.1 Repeated measures ANOVA

Used when one subject receives multiple treatments.

Construct a table similar to that of standard ANOVA: subjects (s_1 to s_n) by treatments (t_1 to t_n).

Again, we need to find within-group (here, within-subject) variance and between-group (between-treatment) variance.

The SS within-subjects is due to variability caused by (a) treatment and (b) individual random variation (termed residual variation).

$$SS_{wit\ subjs} = SS_{treat} + SS_{res}$$

and

$$SS_{res} = SS_{wit\ subjs} - SS_{treat}$$

We calculate the sum of squares within-subjects

$$SS_{wit\ subjs} = \sum_s \sum_t (X_{ts} - \overline{S}_s)^2$$

and for the treatment

$$SS_{treat} = n \sum_t (\overline{T}_t - \overline{X})^2$$

Degrees of freedom:

$$\begin{aligned} DF_{treat} &= m-1 \\ DF_{res} &= DF_{wit\ subjs} - DF_{treat} \\ &= n(m-1) - (m-1) \\ &= (n-1)(m-1) \end{aligned}$$

Mean squares:

$$\begin{aligned} MS_{treat} &= \frac{SS_{treat}}{DF_{treat}} \\ MS_{res} &= \frac{SS_{res}}{DF_{res}} \end{aligned}$$

And finally the F-statistic:

$$F = \frac{MS_{treat}}{MS_{res}} \tag{13.1}$$

13.2 Friedman test

Ordinal (rank) variables. Nonparametric method.

Steps:

- Rank each subject's responses to the treatments, regardless of other subjects
- Sum the ranks for each treatment
- Compute Friedman's statistic, χ_r^2

13.3 Cochrane Q

Chapter 14

Testing for trends

14.1 Regression

See Machine Learning file.

14.2 Correlation

For more detail, see the Machine Learning file.

Covariance:

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Correlation:

$$\begin{aligned} \rho &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \end{aligned}$$

14.2.1 Significance tests

14.3 Relationship between correlation and linear regression

Regression is used to explain a change in Y with respect to X, and it implies causation of X. Correlation only measures association, stating nothing about

causation.

Correlation:

$$\rho = \frac{cov(x,y)}{s_x s_y}$$

Regression:

$$\beta = \frac{cov(x,y)}{s_x^2}$$

So

$$\rho = \beta \frac{s_x}{s_y}$$

and

$$\beta = \rho \frac{s_y}{s_x}$$

The square of the correlation coefficient, r^2 is the **coefficient of determination**, which measures the degree to which a straight line measures the relationship.

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

It is said that r^2 is the fraction of total variance "explained" by the regression equation.

Chapter 15

Frequencies and Goodness of Fit

A goodness-of-fit test is one which tests a hypothesis without an alternative.

Common goodness-of-fit tests are

- Chi-square test
- Kolmogorov test
- Cramer-Smirnov-Von-Mises test
- runs

15.1 Likelihood ratio test

Chapter 16

Multiple Hypothesis Testing

If we perform multiple tests, each with a p-value cutoff of, e.g. 0.05, then 1/20 tests will be falsely positive on average.

16.1 Bonferroni correction

A new p-value cutoff is obtained by dividing the original cutoff by the number of hypotheses.

The Bonferroni inequality:

$$\begin{aligned} \alpha_T &< k\alpha \\ \frac{\alpha_T}{k} &< \alpha \end{aligned} \tag{16.1}$$

where α_T is the true probability of incorrectly rejecting the null at least once (usually the original cutoff).

Notes:

- Extremely conservative. Works well with just a few groups, but as the number of hypotheses becomes large, it becomes more stringent than necessary.
- Assumes independent hypotheses

16.2 Holm test

The Holm test orders the unadjusted p-values and accepts/rejects the tests with decreasing stringency, based on the number of tests already done.

16.3 Tukey

The Tukey statistic q is similar to the t-statistic, but the sampling distribution used for critical value calculation includes a mathematical model of the multiple hypothesis testing problem.

16.4 Student-Newman-Keuls (SNK)

Nonparametric method.

SNK is derived from Tukey, but it is less conservative (finds more differences).

Tukey controls the error for all m comparisons, where SNK only controls for p comparisons under consideration.
(A.k.a. Dunnett's for a single group)

16.5 False Discovery Rate

Chapter 17

Survival Analysis

17.1 Censored Data

Some subjects are not followed to the endpoint (e.g. death). These subjects are "censored".

17.2 Survival curves

Kaplan-Meier survival estimate:

$$\hat{S}(t_j) = \prod_i \frac{n_i - d_i}{n_i} \tag{17.1}$$

where
n_i = num of individuals alive at time t_i
d_i = num of deaths at time t_i

The **median survival time** is the survival time for which the estimated survival function $\hat{S}$ is 0.5.

17.3 Comparing Survival Curves

Note: could use Mann-Whitney or Kruskal-Wallis if data are not censored.

Hazard ratio

The hazard ratio (HR) is equal to the odds = P/(1-P).

For two curves $S_1(t)$ and $S_2(t)$, the hazard ratio ψ is a constant that describes their relationship:

$$S_2(t) = [S_1(t)]^{\psi} \tag{17.2}$$

The hazard function is the probability that a person who survived until time t dies at time t, and is

$$h(t) = \frac{1 - S(t)}{S(t)} \tag{17.3}$$

17.3.1 Log-rank test

Find the expected number of deaths in group 1 at time t_i:

$$e_{1,i} = \frac{n_1 \times d_{total}}{n_{total}}$$

where
n_1 is the number of subjects in group 1
d_{total} is the number of deaths in group 1 and 2
n_{total} is the number of subjects in group 1 and 2

The log-rank test statistic is

$$U_L = \sum_i (d_{1,i} - e_{1,i})$$

which can be normalized and compared to the normal distribution.

17.3.2 Gehan's test

Generalization of Wilcoxan signed rank test

17.3.3 Cox proportional hazards regression

Log-rank and K-M don't work as well for continuous values (e.g. time or gene expression). Cox proportional hazards regression is used for continuous variables. [Although note that it can also be used for categorical data by using dummy 0,1 variables.]

Deals with censored data.

Two survival curves are proportional if the hazard rate h_1 for group 1 is a constant multiple of the hazard rate h_2 for group 2.

Hazard

$$h(t, X) = h_0(t) exp\{\sum_{i=1}^{p} \beta_i X_i\}$$

Hazard ratio: ratio of two hazards

$$\frac{h_1(t)}{h_2(t)}$$

Use ML to estimate parameters (coefficients).

17.3.4 Probability of Hazard

Allows time-dependent covariance, whereas logistic regression requires discretized time.

Non-parametric – important because survival times are often not normally distributed.
(– MW)

17.3.5 Papers

Mike's NEJM paper

Part IV

Parameter Estimation

See Machine Learning file.

Part V

Bayesian Methods

Chapter 18

Bayesian vs Frequentist

Frequentist

- Unknown but fixed parameters
- Estimate parameters with some confidence
- Predict using estimated parameter

Bayesian

- Uncertainty about known parameter
- Use probability to quantify uncertainty: unknown parameters are random variables.
- Predict by rules of probability: Expectation over unknown parameters. Prediction is inference in a Bayes net.