STA248H Condensed Review by David Warde-Farley (dwf at cs dot toronto dot edu)

Preliminaries

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ X \perp Y &\Rightarrow \mathrm{Cov}(X,Y) = 0 \end{aligned}$$

Converse of above implication not necessarily true except if X, Y jointly normally distributed.

Chapter 6

A statistic is a number in context, that summarizes or denotes some property of a set of observed data, usually n i.i.d. realizations of a r.v. X. It is a random variable – possible sources of randomness include measurement error, or the fact that the observed data is a random sample of a larger population, etc.

"5 number summary": min, 1st quartile, median, third quartile, max.

s^2 : unbiased estimator for σ^2.

$$\begin{aligned} s^2 &= \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1} \\ &= \frac{n\sum_{i=1} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n-1} \end{aligned}$$

Population variance, normalize with n rather than $n-1$.

A measure is *robust* if it is not overly affected by outliers (data that is unusually large or small). Examples: median, IQR, mode.

Non-robust measures: mean, standard deviation, range (min/max).

Chapter 7

Weak Law of Large Numbers $X_1, X_2, X_3 ... X_n$ are a sequence of n i.i.d. random variables with $E[X_i] = \mu$ and $\overline{X_n} = \frac{1}{n}\sum_i X_i$, then for any $\epsilon > 0$,

$$P(|\overline{X_n} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

Central Limit Theorem

$$\lim_{n\to\infty} P\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x)$$

Where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution $\mathcal{N}(0,1)$. Consequently, when n is large, $\overline{X}_n$ has approximately normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Point Estimators
An estimator $\hat{\theta}$ for parameter θ is *unbiased* if

$$E[\hat{\theta}] = \theta$$

An estimator $\hat{\theta}$ for parameter θ is *consistent* if

$$\forall \epsilon > 0, \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

Ideally a good estimator should be unbiased, consistent, and have small variance for large sample sizes, since statistical studies typically aren't done over and over again; we only get one estimate and we want it to be relatively close to the true parameter.

Sample mean $\overline{X}$ is unbiased estimator for μ. s^2 is an unbiased estimator σ^2.

$$\begin{aligned} \text{Var } \overline{X} &= \frac{\sigma^2}{n} \\ \text{Std. Error} &= \sqrt{\text{Var } \overline{X}} = \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Some Useful Identities

$$\begin{aligned} E[aX] &= aE[X] \\ E[X+Y] &= E[X] + E[Y] \\ \mathrm{Var}(aX) &= a^2\mathrm{Var}(X) \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \iff \mathrm{Cov}(X,Y) = 0 \end{aligned}$$

Identities for Normal Random Variables

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \sigma_X^2) \\ Y &\sim \mathcal{N}(\mu_Y, \sigma_Y^2) \end{aligned}$$

$$\begin{aligned} E[X+Y] &= \mu_X + \mu_Y \\ \mathrm{Var}(X+Y) &= \sigma_X^2 + \sigma_Y^2 + 2\mathrm{Cov}(X,Y) \\ X+Y &\sim \mathcal{N}(\mu_x + \mu_y, \mathrm{Var}(X+Y)) \end{aligned}$$

$$kX \sim \mathcal{N}(k\mu_X, k^2\sigma_X^2)$$

In fact, any linear combination of normal random variables is normal with some mean and variance.

Method of Moments

$$M_k = E[X^k]$$

Express the parameter to be estimated in terms of its moments, equate the sample moments with the theoretical moments. Solve for $\hat{\theta}$.

Maximum Likelihood

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Where f is probability density function of X, and $x_1, x_2 ... x_n$ are n i.i.d. realizations of X.

Maximize $L(\theta)$ or $\log L(\theta)$, solve for θ, replace with $\hat{\theta}$.

Often easier to optimize the log likelihood, and because log is a monotonically increasing function, taking the log of a function doesn't change where the max is.

Why MLEs more common than MofM estimators

1. MLEs are consistent
2. When n is large, MLEs have variance nearly as small as any estimator can achieve.
3. When n is large, sampling distribution of MLEs is approximately normal.

Measure of an estimator's precision: **standard error** (S.D. of sampling distribution)

$$\begin{aligned} SE(\hat{\theta}) &= \sqrt{\text{Var}(\hat{\theta})} \\ &= \sigma/\sqrt{n} \\ &= s/\sqrt{n} \text{ (estimated std. error)} \end{aligned}$$

Parametric Bootstrap
Know the form of the distribution $f(x|\theta)$ (i.e. normal, binomial, exponential, etc.) and we'd like to evaluate the result of our estimator $\hat{\theta}$, without doing much thinking. So we do computation instead.

1. Estimate $\hat{\theta}$ from data $x_1, x_2 ... x_n$.
2. Simulate a number of "bootstrap samples" of size n from $f(x|\hat{\theta})$ and for each of these calculate an estimate θ the same way you calculated your original estimator. Call these $\hat{\theta}_i^\star$. The number of these "bootstrap estimates" you make is usually called B and is usually 100, 200, 500, 1000, etc.
3. Let $\overline{\theta}^\star = \frac{1}{B}\sum \hat{\theta}_i^\star$ (i.e. the sample mean of all the boostrap estimates).
4. Now calculate the sample standard deviation of the $\hat{\theta}$.

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^\star - \overline{\theta}^\star\right)^2}$$

Voila, there's your bootstrap estimate of the standard error (remember, standard error is the s.d. of the sampling distribution of $\hat{\theta}$, which is what we've simulated to find out here).

Non-parametric Bootstrap
For when we don't really know the form of the distribution $f(x|\theta)$, and don't want to make any strong assumptions about it.

Regard original data as the population. Get B bootstrap samples as before, but this time instead of randomly generating pseudo data, we *sample with replacement from the original data*. Again, each bootstrap sample should be of size n (same as original data). You can then calculate the $\hat{\theta}_i^\star$'s and then take their mean to get $\overline{\theta}^\star$, your "bootstrap estimate" of the parameter. You can then calculate standard error as before or use this in lieu of your $\hat{\theta}$.

Bootstrap estimates for parameters have some nice theoretical properties, including consistency.

Confidence Intervals
Range of plausible values for an estimate of a parameter. A $100(1-\alpha)\%$ will catch the true parameter in between the upper and lower limit that percentage of the time, and miss it the rest of the time just by chance.

Confidence interval for mean of a normal distribution, σ known (unrealistic):

$$\begin{aligned} \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ \therefore\ z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} &\sim \mathcal{N}(0,1) \end{aligned}$$

So for a $100(1-\alpha)\%$ confidence interval need the $\alpha/2$ quantile of the distribution and the $1-\alpha/2$ quantile of the distribution. For $\mathcal{N}(0,1)$ and $\alpha = 0.05$ (95% confidence) this is -1.96 and +1.96. So,

$$P(-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$$

Solve for μ in the center,

$$\begin{aligned} P(-\overline{X} - 1.96\sigma/\sqrt{n} \le &-\mu &\le -\overline{X} + 1.96\sigma/\sqrt{n}) = 0.95 \\ P(\overline{X} + 1.96\sigma/\sqrt{n} \le &\mu &\le \overline{X} - 1.96\sigma/\sqrt{n}) = 0.95 \end{aligned}$$

So our 95% confidence interval for μ is $\overline{X} \pm 1.96\sigma/\sqrt{n}$.

In general for a normal distribution, σ known, $100(1-\alpha)\%$ C.I. for μ given by

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Approximately good for n large and data non-normally distribution (the further the data are from a normal distribution, the larger the n you need; this works because of CLT).

If σ unknown (i.e. usually), but data follow a normal distribution, replace sigma with s, and you get an "approximate" confidence interval with the above (confidence is less than $100(1-\alpha)\%$).

If data are not normally distributed, n small, σ unknown, use t distribution with $n-1$ degrees of freedom instead of $\mathcal{N}(0,1)$. Then C.I. for μ is

$$\overline{X} \pm t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}$$

In general, bigger sample size → narrower C.I. The higher the confidence you want, the wider the C.I. will have to be.

Bootstrap Confidence Intervals
Sometimes we can't calculate the quantiles/percentiles; we can estimate them from bootstrap samples (can use parametric or non-parametric).

Draw $B = 1000$ samples of size n, estimate $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \hat{\theta}_3^\star ... \hat{\theta}_B^\star$. Overall bootstrap estimate same as before:

$$\overline{\theta}^\star = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^\star$$

Now order the differences $\hat{\theta}_i^\star - \overline{\theta}^\star$, the 25th of these will be the 2.5th percentile of the distribution distribution of $\hat{\theta} - \theta$, and the 975th will be the 97.5th percentile. Thus our 95% confidence interval is

$$\left(\hat{\theta} - (\hat{\theta}_{975}^\star - \overline{\theta}^\star), \hat{\theta} - (\hat{\theta}_{25}^\star - \overline{\theta}^\star)\right)$$

where $\hat{\theta}$ is estimated from the original data.

Chapter 8

Hypothesis testing: analogous to proof by contradiction. Assume *null hypothesis* (H_0) is true, look for evidence for alternative hypothesis (H_a). Reject if observed data is improbable under the null hypothesis assumption.

Type I error: rejecting the null hypothesis when it is true.
Type II error: not rejecting the null hypothesis is false.

$$\begin{aligned} P(\text{Type I Error}) &= \alpha \\ P(\text{Type II Error}) &= \beta \end{aligned}$$

Neyman-Pearson model: fix α, usually 0.01, 0.05, 0.1. The "significance level".

Test statistic: Statistic whose distribution is known under the null hypothesis.

For mean of a normal distribution σ known, can use test statistic

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Will have $\mathcal{N}(0,1)$ distribution under the null hypothesis that $\mu = \mu_0$ (also works with s instead of μ if $n > 40$, about). Depending on your H_a, which could be any of $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$ (two-sided test), you would choose your rejection region appropriately.

More realistic test (σ unknown):

$$t_{\text{obs}} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

This will follow a t-distribution with $n-1$ degrees of freedom if H_0 is true, and the data is approximately normally distributed.

The **power** of a test is $1 - \beta$, i.e. the probability that we will correctly reject the null hypothesis when you should. Needs to be evaluated at a certain value of μ', so β is actually a function of the alternative value that borders our rejection region.

Significance testing
Another method of hypothesis testing, gives you an idea of the strength of your results, via a "p-value" – the probability of observing the value of the test statistic that was observed, or a value more extreme, given the assumption that H_0 is true.

$p > .1$	no evidence against H_0
$.05 < p < .1$	weak evidence against H_0
$.01 < p < .05$	moderate evidence against H_0
$p < .01$	strong evidence against H_0

Bootstrap testing
Say we want to test the hypothesis $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$. So $p = P(\overline{X} \geq \mu_0 | H_0 \text{ is true})$.Need to estimate p-value under H_0 assumption, so bootstrap the new data values $y_i = x_i - \overline{X} + \mu_0$, this way $E[Y] = \mu_0$.

Sample with replacement (non-parametric bootstrap) for B samples, as usual, from $y_1, y_2, ...y_n$ and calculate bootstrap means $\overline{Y}_1^\star, \overline{Y}_2^\star, ...\overline{Y}_B^\star$. Bootstrap estimate of p-value is

$$\hat{p}^\star = \frac{\text{Number of the bootstrap means } \overline{Y}_i^\star \text{ that are} \geq \overline{X}}{B}$$

Chapter 9 (9.1 and 9.2)

Estimating proportions: The typical situation, we have a population of interest and a particular trait; some of the population have it, some don't. We want to make inferences on p, the proportion of the population with the trait (not to be confused with the p-value from Chapter 8). We assume our observed data $x_1, x_2, ...x_n$ are binary random variables

$$x_i = \begin{cases} 1, & \text{sample point } i \text{ has the trait} \\ 0, & \text{sample point } i \text{ lacks the trait} \end{cases}$$

Then the proportion of the sample that has the trait (and in fact the unbiased maximum likelihood estimator for p) is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note that this is the sample mean of n observations of our binary random variable. By the Central Limit Theorem, $\hat{p}$ is normally distributed with the same mean as X (which is p, the true proportion that have the trait) and with variance $\text{Var}(X)/n = p(1-p)/n$. To get a confidence interval for this

quantity, we would use $\hat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}$, but that presupposes knowledge of p, which is the quantity we're estimating. so replace p with $\hat{p}$ and we get

$$p \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Since we're using the sample s.d. instead of the real s.d., you might think we should use a t_{n-1} distribution, but remember that for $n > 40$ or so, t distribution is approximately standard normal, so it doesn't make much of a difference in real experimental settings where n is certainly > 40.

We may get a confidence interval so big that its useless, depending on our sample size n. As for exactly how large n should be, for $d = |p - \hat{p}|$ "degree of accuracy", and assuming we have a prior estimate of p,

$$n \doteq \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{d^2}$$

If we don't have a prior estimate of p,

$$n \doteq \frac{z^2_{\alpha/2}}{4d^2}$$

Since it can be shown that $\hat{p}(1-\hat{p})$ will never exceed $1/4$ (the function has a maximum at $p = 1/2$, the point with the most uncertainty, thus the highest variance).

To test a hypothesis on a proportion, i.e. where $H_0 : p = p_0$, use the test statistic

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

If H_0 is true, the above will have a standard normal distribution.

Chapter 10

Comparing two means: Most typically,

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 \neq 0 \end{aligned}$$

If data are normally distributed (don't we always wish they were), the test statistic

$$T = \frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has approximately a t-distribution. For hand calculations use $df = \min(n_1 - 1, n_2 - 1)$ for a conservative estimate.

A common additional assumption is that $\sigma_1^2 = \sigma_2^2$, in which case we should use the pooled estimate

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

and our test statistic becomes

$$T = \frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Confidence interval on the difference in means
To obtain a confidence interval for $\mu_1 - \mu_2$, use

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha_2, df}\, s_{\overline{Y}_1 - \overline{Y}_2}$$

Where df and s differ based on your assumptions.

If you assume variances *are not* equal, use $df = min(n_1 - 1, n_2 - 1)$ or the more complicated version in the notes, and

$$s_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{s^2_{Y_1}}{n_1} + \frac{s^2_{Y_2}}{n_2}}$$

If you assume variances *are* equal, use $df = n_1 + n_2 - 2$ and

$$s_{\overline{Y}_1 - \overline{Y}_2} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Paired data
When data consist of paired observations from two groups, i.e. 15 pairs of identical twins who differ in some significant way, we can't treat them as we normally do independent samples. So we look at differences between the pairs and carry out a t-test or other such thing. So if our H_0 is that there is no difference,

$$t_{\text{obs}} = \frac{\overline{d} - 0}{s_d/\sqrt{n}}$$

has a t distribution with $df = n - 1$ (n is the number of pairs, not the total number of observations!).

Transformations
If there is a strong skew, a log transformation will often reduce the skew of data. Log transformations also "stabilize" the variance, reduce the effects of variances that increase with the mean.

Square root transformations work pretty well for less strongly skewed data.

Two-sample Bootstrap tests
If we've got two samples of $x_1, x_2, ...x_n$ and $y_1, y_2, ...y_n$, of size n_1 and n_2 respectively, and want to test

$$\begin{aligned} H_0 &: \mu_Y - \mu_X = 0 \\ H_a &: \mu_Y - \mu_X > 0 \end{aligned}$$

Our test statistic is $V = \overline{Y} - \overline{X}$. Observed value v.
p-value: $P(V \geq v | H_0 \text{ is true})$.

Want a bootstrap estimate of the p-value. Need to generate bootstrap samples assuming H_0 is true, i.e. they're from the same distribution, and create one big group of size $n_1 + n_2$; resample it with replacement into two new groups of size n_1 and n_2; do this B times and each time calculate $v_j^\star = \overline{y}_j^\star - \overline{x}_j^\star$, $j = 1, 2, ...B$. The bootstrap estimate of p is then

$$\hat{p}^\star = \frac{\text{number of times } v_j^\star \geq v}{B}$$

Data Collection Methods
Three main methods (strength of evidence needed to support conclusions increases down this list)

1. Observational studies:
   - No intervention
   - Problem: **confounding** - can't separate the effects of some variables with others.
2. Sample surveys (subset of observational)
   - random sample chosen from a population in order to make inference on entire population (as opposed to a census where we record for entire population)
   - choose randomly to avoid bias
   - still observational, can still suffer from confounding
3. Experiments
   - Subjects randomly assigned to intervention
   - Actually capture cause-effect conclusions
   - **Interventions:** Predictor variables are called *factors*, values of factors are called *levels*.
   - Each combination of levels of factors is a "treatment"
   - **Key step:** randomization used to eliminate the effects of confounding

Principles of Experimental Design

1. Control - group for comparison
2. Randomization
3. Replication - need multiple observations/treatments

The drawback is that we can't always carry out controlled experiments (i.e. smoking and lung cancer).
Placebos are treatments with no scientific effect.
A **double blind** is used when neither the test subject nor the person administering the test knows which treatment is being used. Eliminates possibility of experimenter bias.

Controlling for Type I error rate
Bonferroni Inequality:

$$P(A_1 \cup A_2 \cup ...A_K) \leq P(A_1) + P(A_2) + ...P(A_k)$$

How does this apply?

$$\begin{aligned} & P(\text{At least one Type I error in } k \text{ tests}) \\ = \; & P(A_1 \cup A_2 \cup ...A_K) \leq k\alpha \end{aligned}$$

So if we use $\frac{\alpha}{k}$ as significance level for each individual test, the "overall significance level" will be *at most* α. This is known as "Bonferroni's method", and it is a conservative procedure – overall significance level is generally $< \alpha$ (i.e. results are as if you tested at a lower significance level → stronger evidence).

Chi-squared distribution and F distribution
Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Let

$$Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Then $Z_i^2 \sim \chi_1^2$, and

$$Z_1^2 + Z_2^2 + ...Z_k^2 \sim \chi_k^2$$

If $X_1 \sim \chi_m^2$ and $X_2 \sim \chi_n^2$, then

$$\frac{X_1/m}{X_2/n} \sim F_{m,n}$$

If $T = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$,

$$T^2 \sim F_{1,n-1}$$

F-test for equality of variance (Section 10.2)
First recall that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Assuming we have two normal distributions with variances σ_1^2 and σ_2^2, our test statistic will be

$$\begin{aligned} & \frac{(n_1-1)s_1^2/\sigma_1^2}{n_1-1} \Big/ \frac{(n_2-1)s_2^2/\sigma_2^2}{n_2-1} \\ = \; & \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \end{aligned}$$

If H_0 is true ($\sigma_1^2 = \sigma_2^2$) then this is s_1^2/s_2^2 and it is a sample from a F_{n_1-1,n_2-1} distribution (easy to see why from the characterization above).

This test is ***very*** sensitive to departures from normality; a significant test might mean $\sigma_1^2 \neq \sigma_2^2$, or it may mean the data aren't from normal distributions, and it's hard to tell which (there are other tests that are less sensitive).

Chapter 13

One-way Analysis of Variance
To be continued... some day...