Notes for 18.6501x, Fundamentals of Statistics
v0.2 (2019 April 24)

Dan Schmidt (edX: `dfannius`)

1 Introduction

These are notes for the spring 2019 session of the MITx class 18.6501x, Fundamentals of Statistics. They are basically just a summary of the slides in complete sentences for my own benefit, with some extra commentary where it was useful to me. The notation does not always match the slides exactly, and there may certainly be errors.

2 Limits for large samples

Law of large numbers: The sample mean approaches the expected value (almost surely) as the sample size goes to infinity.

Central limit theorem:

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0,1). \tag{1}$$

μ and σ are the mean and standard deviation of the actual distribution that X is drawn from.

These two properties underlie everything we do! The law of large numbers means that we can replace expectations with averages when performing estimates, and the central limit theorem means that no matter what we measure, we can end up using the same yardstick to talk about things like rarity or confidence.

Hoeffding inequality: If n is a positive integer and X_i are i.i.d. and are in the range $[a, b]$ almost surely, then

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\}. \tag{2}$$

3 Gaussian distribution

Also known as the normal distribution. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3}$$

$$\mathbb{E}[X] = \mu \tag{4}$$

$$\mathrm{Var}(X) = \sigma^2 \tag{5}$$

We use Z for a standard normal ($\mathcal{N}(0,1)$). $\Phi(z)$ is defined to be $\int_{-\infty}^{z} f(z)\,\mathrm{d}z$, the probability that a random sample for Z will be less than z.

q_α is the number such that $\Phi(z) = 1 - \alpha$. So $q_{0.01}$ will put you at the 99th percentile of a standard normal.

4 Types of convergence

T_n is a sequence of random variables converging to the random variable T. We consider three main types of convergence in descending order of strength.

Almost sure convergence means that the values of T_n converge to T.

$$T_n \xrightarrow[n\to\infty]{\text{a.s}} T \text{ iff } \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1 \tag{6}$$

Convergence in probability means that the chance that drawing from T_n does not yield a number arbitrarily close to T becomes arbitarily small.

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \text{ iff } \mathbb{P}\left[|T_n - T| \geq \epsilon\right] \xrightarrow[n\to\infty]{} 0, \forall \epsilon > 0 \tag{7}$$

Convergence in distribution means that any reasonable function on our T_n is going to yield the same result if we call it on T.

$$T_n \xrightarrow[n\to\infty]{\text{(d)}} T \text{ iff } \mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)], f \text{ bounded and continuous} \tag{8}$$

For almost sure convergence and convergence in probability (but not convergence in distribution),

$$T_n + U_n \xrightarrow[n\to\infty]{} T + U \tag{9}$$

$$T_n U_n \xrightarrow[n\to\infty]{} TU \tag{10}$$

$$\frac{T_n}{U_n} \xrightarrow[n\to\infty]{} \frac{T}{U} \tag{11}$$

4.1 Slutsky's theorem

Convergence in distribution isn't good enough by itself for the above manipulations but is okay if combined with convergence in probability.

If $T_n \xrightarrow[n\to\infty]{\text{(d)}} T$ and $U_n \xrightarrow[n\to\infty]{\mathbb{P}} u$, then

$$T_n + U_n \xrightarrow[n\to\infty]{} T + u \tag{12}$$

$$T_n U_n \xrightarrow[n\to\infty]{} Tu \tag{13}$$

$$\frac{T_n}{U_n} \xrightarrow[n\to\infty]{} \frac{T}{u} \tag{14}$$

We will use this all the time when making subsitutions with estimators in an expression that converges in distribution.

4.2 Continuous mapping theorem

If f is continuous, then for all three types of convergence,

$$T_n \xrightarrow[n\to\infty]{} T \Rightarrow f(T_n) \xrightarrow[n\to\infty]{} f(T). \tag{15}$$

5 Inference

5.1 Definitions

An experiment produces outcomes in the set E, called the **sample space**. A **sample** is a set $X_i \in E$ of i.i.d. random variables from a distribution $\mathbb{P}$. A **statistical model** is the pair $\big(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$. $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a **family of probability measures** on E. Θ is a **parameter set** from which θ can be drawn.

A model is **well specified** if there exists a θ (called the **true parameter**) such that our actual distribution $\mathbb{P} = \mathbb{P}_\theta$.

If $\Theta \subseteq \mathbb{R}^d$ then the model is **parametric**.

The parameter θ is **identifiable** if it is the only value of θ that produces its distribution $\mathbb{P}_\theta$.

5.2 Estimation

A **statistic** is any measurable function of a sample. An **estimator** is a statistic used to estimate a parameter (it cannot explicitly depend on the parameter).

An estimator $\hat{\theta}_n$ of θ is **consistent** if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta. \tag{16}$$

An estimator $\hat{\theta}_n$ of θ is **asymptotically normal** with **asymptotic variance** σ^2 if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \sigma^2). \tag{17}$$

The $\sqrt{n}$ shows up just because that turns out to be the order of convergence of all good estimators (due to the central limit theorem), so we factor it out so that our expression does not depend on n. Note that when you divide by $\sqrt{n}$, the variance on the right-hand side gets a factor of $\frac{1}{n}$, to make the units work out.

The **bias** of an estimator $\hat{\theta}_n$ is $\mathbb{E}[\hat{\theta}_n - \theta]$. An estimator is **unbiased** if its bias is zero. It is perfectly possible for an estimator to be consistent but biased, if it tends to approach the true parameter value from below, say.

The **quadratic risk** of an estimator $\hat{\theta}_n$ is

$$R(\hat{\theta}_n) \equiv \mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] \tag{18}$$

$$= \text{variance} + \text{bias}^2 \tag{19}$$

as can be easily shown.

5.3 Confidence intervals

A **confidence interval** of level $1-\alpha$ for θ is an interval $\mathcal{I}$ such that the probabiility that $\theta \in \mathcal{I}$ is $1-\alpha$. We complement α because we tend to speak of low αs (e.g, 0.05) and high confidence intervals (e.g., 95%). It is an **asympotic confidence interval** if it is a limit as n goes to infinity.

The standard way to compute a confidence interval is to use the central limit theorem:

$$\sqrt{n}\left(\frac{\hat{\theta}_n - \theta}{\sigma}\right) \xrightarrow[n\to\infty]{} \mathcal{N}(0,1). \tag{20}$$

Now that the right-hand side is a standard normal, we can easily construct confidence intervals of the desired level using things like q_α, and then solve for the interval in which θ must lie.

When we compute a confidence interval in this way, it is likely to depend on the actual parameter, usually in the standard deviation in the denominator, which is too bad because the whole point is that we don't know it. There are three main ways to deal with this:

- **Conservative bound:** put a provable conservative upper bound on the standard deviation (e.g., the maximum possible standard deviation for Bern(p) is $\frac{1}{2}$).
- **I-solve:** keep the standard deviation of the parameter in our equations, and then solve explicitly for the parameter (which will also appear in the numerator). This might be a quadratic equation or worse. Then we can find an interval for the parameter.
- **Plug-in:** if our estimator is consistent, then as n increases, it will approach θ, and by Slutsky's theorem, we can substitute it for θ in the calculation of the standard deviation in the denominator.

5.4 Delta method

This all works fine when our natural estimator is directly for θ. But what if our natural estimator is for $\frac{1}{\lambda}$, say, but we want an estimate for λ? Then we need to transform the left-hand side of the central limit theorem.

If g is a function that is continuously differentiable at the point θ, and we have a sequence of approximations Z_n such that

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \sigma^2) \tag{21}$$

then

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \left(\frac{\mathrm{d}g}{\mathrm{d}x}\Big|_{x=\theta}\right)^2 \sigma^2). \tag{22}$$

This is basically just the chain rule, and makes sense intuitively; if g causes our values to spread out by some factor, then the standard deviation of the resulting values will be multiplied by the same factor.

6 Hypothesis testing

Now we're no longer estimating a parameter θ, but instead determining whether it is likely to be in the disjoint sets Θ_0 or Θ_1. The **null hypothesis** H_0 is the hypothesis that $\theta \in \Theta_0$; generally, this is the status quo, and in order to say that our data has indicated something interesting we need to reject the null hypothesis by showing that the likelihood of any $\theta \in \Theta_0$ generating our data is quite low. In general we will not have to look over all of Θ_0 because it will be clear that there is some subset $\Theta_0' \subset \Theta_0$, usually consisting of just a point or two, such that the $\theta \in \Theta_0$ that maximizes the likelihood of our anomalous data is guaranteed to come from Θ_0'.

We have a **one-sample test** if we're just making a hypothesis about the parameters of one distribution, and a **two-sample test** if we're making a hypothesis comparing two distributions (e.g., whose mean is bigger).

Our general plan is to construct the null hypothesis H_0 and then look for evidence to reject it. We make a binary test $\psi(X)$ that returns a 1 if H_0 is to be rejected and 0 otherwise.

The **type 1 error** of ψ, α_ψ, is the probability that we will reject H_0 when the data indeed comes from H_0. The **type 2 error**, β_ψ, is the probability that we will not reject H_0 even though the data comes from H_1. The **power** of a test is one minus the maximum type 2 error. Basically, it is the confidence we can have, once we have not rejected H_0, that H_1 is not true.

A test ψ has **level** α if its type 1 error is at most α. It has **asymptotic level** α if its type 1 error is at most α as n goes to infinity.

In general the test will have the form

$$\psi(X) = \mathbb{1}\{T_n > c\} \tag{23}$$

where T_n, the **test statistic**, is a function of X and c is a constant that determines the **rejection region**.

In a **one-sided test** we will only reject H_0 if our data is extreme in one direction (e.g., if our null hypothesis is that the true mean is at most μ_0, we'll only reject if we see a large mean). In a **two-sided test** we will reject H_0 if our data is extreme in either direction (e.g., if our null hypothesis is that our mean is exactly μ_0).

The general procedure is to use the central limit theorem to construct a test statistic T_n that can be compared to a standard normal. For a one-sided test we will reject if $T_n > q_\alpha$; for a two-sided test we will reject if $|T_n| > q_{\alpha/2}$ (since there are two tails to reject).

In this manner we can construct a family of tests ψ_α that each have level α by adjusting the rejection region. The **p-value** of a sample is the smallest α at which ψ_α will reject H_0 when presented with the sample. Basically, it says "The fraction p of all data generated by H_0 looks like your data or is even more extreme". Thus, a smaller p-value lets us reject H_0 more confidently.

7 Methods of estimation

7.1 Total variation distance

If our estimate of the parameter(s) is good then our estimated distribution will look a lot like the true distribution. How can we quantify this?

The **total variation distance** between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is the event A (this is any subset of the sample space!) that maximizes the difference between $\mathbb{P}_\theta(A)$ and $\mathbb{P}_{\theta'}(A)$. Intuitively, this is all of the points x for which $P_\theta(x) > P_{\theta'}(x)$. Since both distributions must integrate to 1, the total variation in the $<$ direction is the same, so

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int |f_\theta(x) - f_{\theta'}(x)| \, dx. \tag{24}$$

The total variation distance is a true distance; it's symmetric, definite (if it's zero, the two distributions are the same), and satisfies the triangle inequality.

7.2 Estimating by minimizing distance

If we had a nice way to estimate the distance (total variation or otherwise) between P_θ and P_{θ^*}, then we could estimate θ^* by optimizing that distance over θ, trying to drive it to zero. That's tough with total variation distance, but there are other expressions that are easier to deal with. For example:

7.3 Kullback-Leibler divergence

This is defined as

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \int f_\theta(x) \log \frac{f_\theta(x)}{f_{\theta'}(x)} dx \tag{25}$$

This is not a distance (it's a divergence) because it's not symmetric, but it is definite.

Since I first encountered KL divergence in the context of information theory, I think of it as "how much do we lose by encoding samples from P_θ with a code based on $P_{\theta'}$?"

7.4 Maximum likelihood estimation

One can show that minimizing the KL divergence is equivalent to maximizing the likelihood of the sample (or, what is usually easier, minimizing its negative log-likelihood). Of course maximizing the likelihood of the observed data makes intuitive sense anyway.

To do so, we just compute the likelihood of our whole sample, which is probably a big product if the sample is i.i.d., take the negative log, which is probably a big sum, take the derivative with respect to the parameters, and set it to 0, then read out our estimates of the parameters. If we are not being sloppy we need to ensure that this is a minimum by looking at the second derivative or the Hessian.

7.5 Concave and convex functions

Jensen's inequality: if $f(x)$ is convex, then $\mathbb{E}(f(x) \geq f(E(x))$.

A function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is concave if its Hessian is negative semidefinite ($\mathbf{x}^{\mathrm{T}}\mathbf{H}\mathbf{x} \leq 0 \, \forall \mathbf{x} \neq \mathbf{0}$) and strictly concave if its Hessian is negative definite.

8 Multivariable estimation

Now each observation consists of multiple variables, e.g., the height and weight of a single person.

8.1 Covariances

$\mathrm{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)\right)$. In fact we only need to center one variable: $\mathrm{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))Y\right)$. By definition, $\mathrm{Cov}(X, Y) = \mathrm{Var}(X)$.

If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$. The converse is not true in general (what if $Y = 0$?) but it is if (X, Y) is a Gaussian vector, meaning that $\alpha X + \beta Y$ is Gaussian for (α, β) except $(0, 0)$.

8.2 Covariance matrix

The covariance matrix of X, annoyingly called Σ instead of Σ^2, is defined by $\Sigma_{ij} \equiv \mathrm{Cov}(X^{(i)}, X^{(j)})$. Even more annoying, this usually is named $\mathrm{Cov}(X)$, not to be confused with $\mathrm{Cov}(X, Y)$. I will try to note the difference by writing Cov when I am discussing a covariance matrix.

The equivalent to $\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X)$ is $\mathsf{Cov}(\mathbf{A}X + \mathbf{b}) = \mathbf{A} \, \mathsf{Cov}(X)\mathbf{A}^{\mathrm{T}}$.

8.3 Multivariate analogies to univariate topics

Gaussian distribution:

$$f(\mathbf{x}) = \frac{1}{(2\pi \det \Sigma)^{d/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{26}$$

Central limit theorem:

$$\sqrt{n}(\overline{X}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{27}$$

$$\sqrt{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\overline{X}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{28}$$

Delta method:

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \boldsymbol{\gamma}^{\mathrm{T}}\Sigma\boldsymbol{\gamma}) \qquad \text{where } \boldsymbol{\gamma} \equiv \left.\frac{\mathrm{d}g}{\mathrm{d}\mathbf{x}}\right|_{\mathbf{x}=\theta}. \tag{29}$$

8.4 Fisher information

There are multiple ways to derive or interpret the Fisher information matrix. The most direct way to interpret it is that its inverse is the asymptotic covariance matrix of the maximum likelihood estimator of the parameters, so the bigger it is, the quicker our estimate will converge.

We define $\ell_X(\theta)$ as the log likelihood of one observation X:

$$\ell_X(\theta) = \log L_1(X, \theta) \tag{30}$$

The Fisher information can be computed in two ways: as the covariance matrix of the gradient of ℓ, and as the negative expected value of the Hessian of ℓ. In both cases we integrate out X and are left with a function of θ only. The fact that we're looking at probability distributions and thus various integrals are 1 (and their derivatives are 0) is key to proving this equivalence.

$$\mathcal{I}(\theta) = \mathbb{E}[\mathbf{g}\mathbf{g}^{\mathrm{T}}] - \mathbb{E}[\mathbf{g}]\,\mathbb{E}[\mathbf{g}^{\mathrm{T}}] \qquad \text{where } \mathbf{g} \equiv \frac{\mathrm{d}\ell}{\mathrm{d}\mathbf{x}} \tag{31}$$

$$\mathcal{I}(\theta) = -\,\mathbb{E}[\mathbf{H}(\ell(\theta))] \tag{32}$$

All of these expectations are over $X \sim L_1(X, \theta)$.

With one variable, the Fisher information is just a number and the formulas are simply

$$\mathcal{I}(\theta) = \mathrm{Var}(\ell'(\theta)) = -\,\mathbb{E}[\ell''(\theta)]. \tag{33}$$

We can show that if $\hat{\theta}$ is the maximum likelihood estimator for θ, with true parameter θ^*, then

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}). \tag{34}$$

8.5 The method of moments

Another way to estimate parameters is to use the moment generating function to write out as many equations for moments as we have parameters. We then have a system of d equations in d unknowns, and given a sample we can find the **empirical moments** (moments of the sample), substitute them into the equations, and solve for our parameter estimates. This works (in the limit) by the law of large numbers.

If we want, we don't have to use the first d moments, but can use any d statistics that are easily calculable and independent.

Let Σ be the $d \times d$ covariance matrix of these moments. Let $M(\theta)$ be the vector of the population moments, so $M^{-1}(\mathbf{m})$ deduces the parameters from the moments. Defining

$$\boldsymbol{\gamma} \equiv \frac{\mathrm{d}M^{-1}}{\mathrm{d}m}\Big|_{m=M(\theta)}, \tag{35}$$

then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \boldsymbol{\gamma}^{\mathrm{T}}\Sigma\boldsymbol{\gamma}), \tag{36}$$

analogously to the delta method.

8.6 M-estimation

Instead of maximizing the likelihood, we could optimize any function ρ of our data and our unknown parameters. Let

$$\mathcal{Q}(\mu) \equiv \mathbb{E}[\rho(X, \mu)]. \tag{37}$$

$\mathcal{Q}$ is a cost function, so we want $\arg\min(\mathcal{Q}) = \mu^*$, the true parameter. This is true if ρ is negative log likelihood, for example, but we can generalize it to any function that is best when it is at a minimum. For example, if $\rho(x, \mu) = |x - \mu|$, then μ^* is the median of the distribution. So with a properly chosen ρ we can estimate any statistic.

We estimate in the usual way by replacing expectations with averages. Let $\hat{\mu}_n$ be the minimizer of

$$\mathcal{Q}_n(\mu) \equiv \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\mu). \tag{38}$$

Then

$$\sqrt{n}(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0,\Sigma) \tag{39}$$

where

$$\Sigma \equiv J(\mu^*)^{-1}K(\mu^*)J(\mu^*)^{-1} \tag{40}$$

$$J(\mu) \equiv \frac{\partial^2}{\partial\mu\partial\mu^{\mathrm{T}}}\mathcal{Q}(\mu) = \mathbb{E}_X\left[\frac{\partial^2}{\partial\mu\partial\mu^{\mathrm{T}}}\rho(X,\mu)\right] \quad \text{(usually)} \tag{41}$$

$$K(\mu) = \mathsf{Cov}_X\left[\frac{\partial}{\partial\mu}\rho(X,\mu)\right]. \tag{42}$$

That's a lot uglier than the MLE calculation, for which J and K were both conveniently equal to the Fisher information so two of them canceled out. In fact it's ugly enough that we have never been asked to do it in the course.

M-estimators can be more robust than some other estimators; the median is a good example.

9 Non-asymptotic hypothesis testing

We've been using the central limit theorem and assuming n is large. But what if n is small? If our samples come from a Gaussian, we're in luck; the central limit theorem formula is exact even for small n. We still have a variance to estimate, though.

9.1 The χ^2 distribution

χ^2_d (the **chi-square distribution with d degrees of freedom**) is distributed like $Z_1^2 + \cdots + Z_d^2$, where Z_i are all independent standard normals. Its expected value is d and its variance is $2d$.

We encounter it in practice when reasoning about the sample variance

$$S_n = \frac{1}{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\left(X_i^2 - \left(\overline{X}_n\right)^2\right). \tag{43}$$

Cochran's theorem tells us that if $X \sim \mathcal{N}(\mu,\sigma^2)$, then $\overline{X}_n$ and S_n are independent, and

$$\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}. \tag{44}$$

The reason we end up with $n-1$ instead of n is that we basically took away one degree of freedom by recentering everything before squaring it. To work backwards, if we fix $\overline{X}_n$, then X_n is completely determined by $X_1,\ldots,X_{n-1}$.

If instead of S_n we use the unbiased estimator of σ^2,

$$\tilde{S}_n = \frac{1}{n-1}\left(X_i - \overline{X}_n\right)^2, \tag{45}$$

then we have

$$\frac{(n-1)\tilde{S}_n}{\sigma^2} \sim \chi^2_{n-1}. \tag{46}$$

9.2 Student's t-distribution

Student's t-distribution with d degrees of freedom (t_d) is $Z/\sqrt{V/d}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$, and Z and V are independent (which they will be when we use it thanks to Cochran's theorem).

Here's how we use it. We have a small number of $X \sim \mathcal{N}(\mu, \sigma^2)$ and are testing the hypothesis $\mu = \mu_0$. We use the test statistic

$$T_n = \frac{\overline{X}_n - \mu_0}{\sqrt{\frac{\tilde{S}_n}{n}}} = \frac{\sqrt{n}\left(\frac{\overline{X}_n - \mu_0}{\sigma}\right)}{\sqrt{\frac{\tilde{S}_n}{\sigma^2}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}, \tag{47}$$

using the exact central limit theorem in the numerator and Cochran's theorem in the denominator. This is in exactly the form of Student's t-distribution, so we can compare T_n to a Student's t quantile just like we compared to Gaussian quantiles when we did large-n hypothesis testing.

Note again that that X has to come from a Gaussian distribution!

9.3 Welch-Satterthwaite formula

That was for a one-sample test. What if we're doing a two-sample test with samples of different sizes? We'd like to find the distribution of

$$\frac{\left(\overline{X}_n - \overline{Y}_m\right) - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}}. \tag{48}$$

This turns out to be t_N, where

$$N = \frac{\left(\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}\right)^2}{\frac{\hat{\sigma}_x^4}{n^2(n-1)} + \frac{\hat{\sigma}_y^4}{m^2(m-1)}}. \tag{49}$$

This is the **Welch-Satterthwaite formula**.

9.4 Wald's test

We have a sample with MLE $\hat{\theta} \in \mathbb{R}^d$ and a hypothesis that the true set of parameters θ^* is equal to some fixed guess θ_0. If H_0 is true, then

$$\sqrt{n}\,\mathcal{I}(\theta_0)^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \mathbf{I}). \tag{50}$$

But if the hypothesis is true, then it is just as true that

$$\sqrt{n}\,\mathcal{I}(\hat{\theta}_n)^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, \mathbf{I}), \tag{51}$$

since $\hat{\theta}_n$ will approach $\theta_0 = \theta^*$ as n increases. Taking the inner product of both sides, we get

$$n\left(\hat{\theta}_n - \theta_0\right)^{\mathrm{T}} \mathcal{I}(\hat{\theta}_n)\left(\hat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\text{(d)}} \chi^2_d. \tag{52}$$

This is **Wald's test**. It's telling us the squared distance of $\hat{\theta}_n$ to θ_0 using $\mathcal{I}(\hat{\theta}_n)$ as our metric. Along dimensions with high information (and low variance), we require the vectors' components to be closer.

9.5 Likelihood ratio test

Here we have a parameter space $\Theta \subseteq \mathbb{R}^d$, and our null hypothesis is that parameters θ_{r+1} through θ_d have certain values θ^c_{r+1} through θ^c_d, leaving the other r unspecified. We construct two estimators

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta) \tag{53}$$

$$\hat{\theta}^c_n = \arg\max_{\theta \in \Theta^c} \ell_n(\theta) \tag{54}$$

Our test statistic is

$$T_n = 2\left(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}^c_n)\right). \tag{55}$$

Wilks' theorem tells us that if H_0 is true and standard conditions are satisfied, then

$$T_n \xrightarrow[n\to\infty]{\text{(d)}} \chi^2_{d-r}. \tag{56}$$

The **likelihood ratio test** consists of comparing this to a quantile of the chi-square distribution.

9.6 Implicit hypotheses

Now our hypotheses depend on some function of our parameters (e.g., H_0 might be $g(\theta) = \theta_1 - \theta_2 = 0$). If we have an estimator for θ, we can use the delta method. Define

$$\Gamma(\theta) \equiv \boldsymbol{\gamma}^{\mathrm{T}} \Sigma(\theta) \boldsymbol{\gamma} \tag{57}$$

$$\boldsymbol{\gamma} \equiv \nabla g(x)\Big|_{x=\theta} \tag{58}$$

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}_d(0, \Sigma(\theta)) \tag{59}$$

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}_k(0, \Gamma(\theta)) \tag{60}$$

$$\sqrt{n}\Gamma(\theta)^{-\frac{1}{2}}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}_k(0, \mathbf{I}_k) \tag{61}$$

where g produces a vector of size k.

9.6.1 Wald's test

We can use Wald's trick again. By Slutsky we can substitute $\hat{\theta}_n$ for θ so

$$\sqrt{n}\Gamma(\hat{\theta}_n)^{-\frac{1}{2}}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}_k(0, \mathbf{I}_k) \tag{62}$$

so if the hypothesis is true and $g(\theta) = 0$,

$$n g(\hat{\theta}_n)^{\mathrm{T}} \Gamma^{-1} g(\hat{\theta}_n) \xrightarrow[n\to\infty]{\text{(d)}} \chi^2_k, \tag{63}$$

and we can do a chi-square test on it.

10 Non-parametric hypothesis testing

Now we don't even know what model we should fit to the data; our hypothesis will not just be, say, picking μ and σ^2 but the fact that the distribution is Gaussian in the first place.

10.1 Goodness of fit testing on a discrete distribution

Say we have a bunch of of multinoulli observations and want to test a hypothesis that they came from $\mathbf{p} = \mathbf{p}^0$. We have the constraint for any $\mathbf{p}$ that $\sum_i p_i = 1$, so there's one fewer degree of freedom than we might think (e.g., a Bernoulli distribution has only one parameter, not two).

The likelihood of any sequence with total counts $(N_1, \ldots, N_k)$ is $\prod_i p_i^{N_i}$. If we incorporate the constraint we find that the MLE estimate $\hat{\mathbf{p}}_i = \frac{N_i}{n}$.

10.1.1 χ^2 test

We would like to use Wald's test and say

$$\sqrt{n}\mathcal{I}\left(\mathbf{p}^0\right)^{1/2}\left(\hat{\mathbf{p}} - \mathbf{p}^0\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K) \tag{64}$$

(that $(\mathbf{p}^0)^{1/2}$ came from the Fisher information matrix).

But because of our constraint on $\mathbf{p}$, the normal distribution on the right really only has $K-1$ degrees of freedom; our Fisher information matrix on the left does not have full rank. So the Gaussian distribution $\hat{\mathbf{p}} - \mathbf{p}^0$ has only $K-1$ dimensions and when we square the magnitude of that difference for Wald's test, it will be distributed as χ^2_{K-1}, not χ^2_K. Calculating the Fisher information matrix for the multinoulli model, we end up with

$$n\sum_{i=1}^{K}\frac{\left(\hat{\mathbf{p}}_i - \mathbf{p}_i^0\right)^2}{\mathbf{p}_i^0} \xrightarrow[n\to\infty]{\text{(d)}} \chi^2_{K-1}. \tag{65}$$

10.2 Goodness of fit testing on continuous distributions

If we are comparing arbitrary distributions, the best way to put them on equal footing is to compare their CDFs. The **empirical CDF** is piecewise constant, with steps at the values of our observations. The law of large numbers tells us that for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{\text{a.s}} F(t), \tag{66}$$

and the **Glivenko-Cantelli theorem**, also known as the **fundamental theorem of statistics**, tells us that

$$\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| \xrightarrow[n\to\infty]{\text{a.s}} 0. \tag{67}$$

Now, $F(t)$ is just the probability that any given observation will be less than t, so the indicator for whether an observation is indeed less than t is a Bernoulli variable with parameter $F(t)$ and therefore has variance $F(t)(1-F(t))$. So when we use the central limit theorem we get that

$$\sqrt{n}\left(F_n(t) - F(t)\right) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0, F(t)(1-F(t))). \tag{68}$$

Donsker's theorem tells us something even stronger, which is that if F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| \xrightarrow[n\to\infty]{\text{(d)}} \sup_{x\in[0,1]}|\mathbb{B}(x)|, \tag{69}$$

where $\mathbb{B}$ is the distribution of **Brownian bridges** on $[0,1]$, which is outside the scope of this course; the important thing is that it's known and you can look up its quantiles.

10.2.1 Kolmogorov-Smirnov test

Now on to actual hypothesis tests. If we hypothesize a particular distribution with CDF $F^0(t)$, then we can use

$$T_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F^0(t)| \tag{70}$$

as our test statistic. This is the **Kolmogorov-Smirnov test**.

How do we compute this test statistic? Well, if our samples are ordered from lowest to highest, then the CDF evaluated at those samples will go linearly from 0 to 1. Since CDFs are monotonic, we'll be maximally different from the hypothesized distribution at the observation points, so we have

$$T_n = \sqrt{n} \max_i \left(\max_{\delta \in \{-1,0\}} \left| \frac{i+\delta}{n} - F^0(X_{(i)}) \right| \right) \tag{71}$$

It is nice to do this all in the $F(X)$ domain rather than the X domain. Let $U_i = F^0(X_i)$; we want this to be distributed uniformly, so if we define G to be its empirical CDF, we get

$$T_n = \sqrt{n} \sup_{x \in [0,1]} |G(x) - x|. \tag{72}$$

Basically, we are evaluating all the same points but doing it by iterating along the y axis instead of the x axis.

We can generate the pivotal distribution on the right-hand side of this by simulation by computing example T_n values a bunch of times, then make a table of quantiles of it. My understanding is that this is the same thing as calculating $\sup_x |\mathbb{B}(x)|$ but maybe more tractable.

10.2.2 Other goodness of fit tests

We just looked at the Kolmogorov-Smirnov test, which used

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \tag{73}$$

as the definition of distance between functions. This is an L1 metric. We could also use an L2 metric, which is the **Cramér-Von Mises test**:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dF(t) \tag{74}$$

or a modified version of that, the **Anderson-Darling test**:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t). \tag{75}$$

10.2.3 Kolmogorov-Lilliefors test

Say we'd like to know if our data is any Gaussian, not specifically $\mathcal{N}(0, 1)$. We can't just plug in our sample mean and sample variance for F^0, because we've made our hypothesis depend on our data, which is always going to make it look better. If we're going to make our hypothesis depend on our sample like that, we'll have to use a different distribution on the right, which gives us the **Kolmogorov-Lilliefors test**:

$$T_n = \sup_{t \in \mathbb{R}} \left| F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t) \right| \tag{76}$$

and look up the result in the Kolmogorov-Lilliefors table.

10.3 Quantile-quantile tests

We can make a visual comparison of two CDFs F and F_n by comparing the values at their quantiles. So we make a plot where the x values of the points are, say, the 100 percentiles of F and the y values are the 100 percentiles of F_n. The closer they are to the same distribution, the closer the points should be to the line $x = y$.

If the y values are larger in magnitude than the x values (the slope of the plot is steep), that means that the y distribution's tails are fatter than those of x; the values have spread out farther at the same quantile. Conversely, if the ys are small and the slope is shallow, the y distribution has thinner tails than those of x.

11 Bayesian statistics

In frequentist statistics, which we have been studying up to now, the parameter set θ is fixed but unknown. In the Bayesian approach, we treat θ as a random variable with some distribution (the **prior**), and update the distribution to the **posterior** based on the observed data using Bayes' rule:

$$\pi(\theta|X_{1\cdots N}) = \frac{L(X_{1\cdots N}|\theta)\pi(\theta)}{L(X_{1\cdots N})}. \tag{77}$$

(The way I always remember this is to start with $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B) = \mathbb{P}(B|A)\,\mathbb{P}(A)$ and then move $\mathbb{P}(B)$ to the denominator on the other side.)

Since the denominator on the right does not depend on θ, we can write this as

$$\pi(\theta|X_{1\cdots N}) \propto L(X_{1\cdots N}|\theta)\pi(\theta), \tag{78}$$

which is particularly useful since the denominator is often hard to compute.

Given the model of L, we can often find a model for $\pi(\theta)$ such that $\pi(\theta|X_{1\cdots N})$ has the same form, making it particularly easy to perform the update calculations in closed form. When this is true, the family of distributions specified by π is called the **conjugate prior** of the family specified by L. For example, the Beta distribution is the conjugate prior of the Bernoulli distribution.

11.1 Non-informative priors

The Bayesian approach requires us to specify a prior on θ specifying our belief about the relative probabilities of different values for it. If we have no particular belief, or are offended by the concept of making up a prior, we can use a **noninformative prior** that treats all possibilities as equally as possible. This can even be **improper** (it does not integrate to 1), because we can still crank through all the machinery and end up with a posterior that makes sense and will be proper as long as we can normalize it.

11.1.1 Jeffreys prior

What should we use for a maximally uninformative prior? It shouldn't depend on exactly how we specify the parameter. If our model for L is Bern(p) and our prior for p is uniform, we'll get a different posterior distribution from the one we would get if our model for L were Bern(p^2) and our prior for p were uniform, even though you'd think that the uniform prior was maximally uninformative in some sense. The **Jeffreys** prior is invariant to such reparameterizations, and is defined as

$$\pi_J(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}. \tag{79}$$

11.2 Bayesian confidence regions

In the Bayesian framework, we have an entire posterior distribution for $\pi(\theta|X_{1\cdots N})$, not just a point estimate such the MLE. So we can get all sorts of statistics from it such as the mean, the median, or the mode (the **maximum a posteriori** or MAP estimate, analogous to the MLE estimate), and can also do things like find intervals with a specified fraction of the probability mass.