Contents

1 Introduction

1.1 Probability distributions

Name	Distribution	Mean	Variance	MGF	Notes
Normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2		
Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$		
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$(1-\frac{t}{\beta})^{-\alpha}$	
Beta	$\frac{\Gamma(\alpha)+\Gamma(\beta)}{\Gamma(\alpha+\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{\alpha+\beta-2}$		
Binomial	$\binom{n}{k}p^k(1-p)^{n-k}$	pn	$p(1-p)n$		$\to N(pn, p(1-p)n)$
Poisson	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ		

The moment generating function is defined as $\mathbb{E}(e^{tX}) = \int e^t x \, d\mu$. Here is a sketch of uniqueness.

Theorem 1.1 (Billingsley, Theorem 30.1)**:** 1. Let μ be a probability measure with moments of all orders:

$$\alpha_k = \int_{-\infty}^{\infty} x^k \, d\mu.$$

If $\sum_{k\geq 0} \frac{\alpha_k r^k}{k!}$ has a positive radius of convergence, then μ is the only probability measure with those moments.

2. If the MGF of μ is g and g is analytic at 0, then μ is the unique probability measure with MGF g.

Proof sketch. To see (1) from (2), note $g^{(k)}(0) = \alpha_k$.

1. We have uniqueness of characteristic functions (if we took $\mathbb{E}(e^{itX})$ instead) basically by the theorem for Fourier transforms.

2. Let $\beta_k = \int_{-\infty}^{\infty} |x|^k \, d\mu$ and suppose the series converges at radius r. Show that $\frac{\beta_k r^k}{k!} = 0$.

3. We want to show the moments determine the characteristic function. Letting φ be the characteristic function φ of μ, we would like it to be true that

$$\varphi(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \int_{-\infty}^{\infty} (ix)^k e^{itx} \, d\mu, \qquad |h| \le r.$$

   as the sum comes from just expanding e^{itx} in power series. Show the partial sums converge by bounding by β_n's.

   For $\varphi^{(k)}(0)$ give the moments, and we get uniquenes son an interval $(-r, r)$. Repeat near the boundary of the interval.

□

Like characteristic functions (Fourier transforms), MGF's turn convolution (addition of random variables) into multiplication.

Some notes about the distributions.

1. The normal distribution is ubiquitous because by the central limit theorem, the sum of many (nice) iid random variables is close to normal.

2. The exponential distribution is memoryless.

3. The sum of gamma random variables (α_n, β) is $(\sum \alpha_n, \beta)$ because MGF's turn addition of rv's into multiplication.

4. The Poisson distribution can be derived from the following axioms. It is the probability that k events occur in an interval of length 1 when

   (a) events occurring in disjoint intervals are independent and

   (b) equal for intervals of the same length.

   (c) the probability of an event occurring I of length $\to 0$ goes $\to 0$.

   The probability of k events in an interval of length 1 is then

$$\lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k e^{-n}}{k!}.$$

1.2 Statistical distributions

Here are distributions common in statistics. They'll be explained later.

Name	is distribution of...	Notes	
$\chi^2_n = \Gamma(\frac{k}{2}, \frac{1}{2})$	$\sum_i Z_i^2$, $Z_i \sim N(0,1)$		
t_n	$\frac{X}{\sqrt{\frac{1}{n}S}}$, $X \sim N(0,1)$, $S \sim \chi^2_n$		
$F_{m,n}$	$\frac{A/m}{B/n}$, $A \sim \chi^2_m$, $B \sim \chi^2_m$		

1.3 Multidimensional normal distribution

For $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, if x is iid standard normal, then Ax is defined to have distribution $N(0, AA^T)$. Note that if the covariance of x is D ($D = I_n$ here), then the covariance of Ax is ADA^T, so the covariance.

Why is this well-defined? (I.e., if $AA^T = BB^T$ then Ax, Bx are identically distributed.) Assume A is full rank (else we need to restrict to a subspace) We calculate the distribution. By change of variables it's

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\det A|} e^{-\frac{|A^{-1}x|^2}{2}}\, dx = \frac{1}{(2\pi)^{\frac{n}{2}}|\det A|} e^{-\frac{x^T(AA^T)^{-1}x}{2}}\, dx.$$

Define the shifted distribution $N(\mu, A), \mu \in \mathbb{R}^n$ in the obvious way.

1.4 Statistics

In statistics, we want to estimate some function[1] of a distribution F[2] parameterized by $\theta' \in \Theta$, given that $F \in \mathcal{F}$ (some family), when we observe samples drawn from F.

1. If $\mathcal{F}$ (i.e., Θ) is finite-dimensional, we're doing parametric statistics. (For example, we assume F is normal—normal distributions are parametrized by $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{\geq 0}$.)
2. If $\mathcal{F}$ is infinite-dimensional, we're doing nonparametric statistics.

The two kinds of statistics are[3]

1. Bayesian: We assume a distribution on priors. Roughly speaking, given observed event (ex. the sample drawn) B, the likelihood that it came from distribution $F \in \mathcal{F}$ is

$$\mathbb{P}(F|B) = \frac{\mathbb{P}(F)\mathbb{P}(B|F)}{\mathbb{P}(B)}.$$

   Our best guess for F is $\arg\max_F \mathbb{P}(F|B)$. We can calculate the expected value for $\theta(F)$; it will be $\sum_F \mathbb{P}(F|B)\theta(F)$.

[1] "statistic." For instance, mean, standard deviation

[2] F refers to cdf, f refers to pdf

[3] for simplicity, suppose we're dealing with $\mathcal{F}$ finite; for $\mathcal{F}$ infinite (as is usually the case), replace $\mathbb{P}$ with probability density and $\sum$ with $\int$.

2. Frequentist: We assume no distribution on priors $\mathbb{P}(F)$ is known. In this case we simply maximize
$$\mathbb{P}(B|F).$$

Frequentist vs. Bayesian axioms:

	Frequentist	Bayesian
Probability is	limiting relative frequency.	degree of belief.
Parameters	are fixed unknown constants.	can be talk about probabilistically.
Statistical procedures	good long run frequency properties.	involve a distribution on θ.

"To combine prior beliefs with data in a principled way, use Bayesian inference. To construct procedures with guaranteed long run performance, such as confidence intervals, use frequentist methods. Generally, Bayesian methods run into problems when the parameter space is high dimensional."

2 Basic inference

Suppose $(X_1, \ldots, X_n)$ is the observed sample, and our estimate for the statistic θ is $\widehat{\theta_n} = g(X_1, \ldots, X_n)$. (Example: θ is one component of θ'. For example, $\theta' = (\mu, \sigma)$ and θ is just μ or σ.)[4]

2.1 Estimators

2.1.1 Error

Definition 2.1: Define the **standard error** by
$$\text{se} = \sqrt{\text{Var}_\theta(\widehat{\theta_n})}.$$

Note we can't find this directly because we don't know the actual distribution F so don't know $\theta(F)$. Suppose $\text{se} = s(\theta)$. The **estimated standard error** comes from estimating F (i.e., θ) first and plugging that value of θ into the formula for se: $\widehat{\text{se}} = s(\widehat{\theta_n})$.

Define **bias** by[5]
$$\text{bias}(\widehat{\theta_n}) = \mathbb{E}_\theta(\widehat{\theta_n}) - \theta.$$
(Note we must be given the actual value of θ to calculate the bias, so this is a function of θ.)

The **mean standard error (MSE)** is
$$\begin{aligned}\mathbb{E}_\theta[(\widehat{\theta_n} - \theta)^2] &= (\overline{\widehat{\theta_n}} - \theta)^2 + \mathbb{E}_\theta(\widehat{\theta_n} - \overline{\theta_n})^2 \\ &= \text{bias}(\widehat{\theta_n})^2 + \text{Var}_\theta(\widehat{\theta_n})\end{aligned}$$

Bias measures how much the average estimate is from the actual value, the second part measures how much the estimate is from the average estimate.

[4]This is suboptimal notation, but θ is used for both in the whole parameter and a function of it in the literature, and that's confusing.

[5]Warning: $\mathbb{E}_\theta$ means average given θ, not over θ.

Example 2.2: 1. Bernoulli(p). $\text{se} = \sqrt{\frac{p(1-p)}{n}}$ and $\widehat{\text{se}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ where $\widehat{p} = \frac{1}{n}\sum_i X_i$.

2. $N(\mu, \sigma)$. $\text{se} = \sqrt{\frac{n-1}{n}}\sigma$ and $\widehat{\text{se}} = \sqrt{\frac{n-1}{n}}\hat{\sigma} = \sqrt{\frac{n-1}{n}}\sqrt{\frac{n}{n-1}}\sqrt{\mathbb{E}\text{Var}_\theta\{X_1, \ldots, X_n\}} = \sqrt{\mathbb{E}\text{Var}_\theta\{X_1, \ldots, X_n\}} = \frac{\sum(X_i - \overline{X})^2}{n-1}$.

   If we knew se, then the distribution of $\widehat{\mu}$ is

$$\frac{\widehat{\mu} - \mu}{\text{se}} \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

   However, if we use $\widehat{\text{se}}$ instead of se we get a t-distribution rather than a normal distribution. For n large, the t-distribution is approximately the same.

2.1.2 Properties

What properties do we want for an estimator?

1. Unbiased: For every $F \in \mathcal{F}$, $\mathbb{E}_{X_1,\ldots,X_n \sim \mathcal{F}}\widehat{\theta_n} = \theta$. (This is actually not so important!)
2. Consistent: $\widehat{\theta_n} \xrightarrow{P} \theta$ as $n \to \infty$.
3. Asymptotically normal (stronger than consistent): $\frac{\widehat{\theta_n} - \theta}{\text{se}} \rightsquigarrow N(0, 1)$. Warning: sometimes we care about this quantity with $\widehat{\text{se}}$!
4. Asymptotic optimal/efficient: among all well-behaved estimators, the MLE has smallest variance.

Exercise 2.3: Explain why the sample variation is given by $\hat{\sigma}^2 = \frac{\sum(X_i - \overline{X})^2}{N-1}$.

Solution. This is an unbiased estimator, while $\frac{\sum(X_i - \overline{X})^2}{N}$ is a biased estimator.

For simplicity, consider the discrete case. Let $(X_i)_{i=1}^n$ denote the sample. We have (see exercise below)

$$\begin{aligned}
\mathbb{E}\text{Var}((X_i)) &= \mathbb{E}(X_i^2) - (\mathbb{E}(X_i))^2 \\
&= \frac{\sum X_i^2}{n} - \frac{\sum X_i^2}{n^2} - \frac{\sum_{i \neq j} X_i X_j}{n^2} \\
&= \frac{n-1}{n}\left(\frac{\sum X_i^2}{n} - \frac{\sum_{i \neq j} X_i X_j}{n(n-1)}\right) \\
&= \frac{n-1}{n}(\mathbb{E}(X_i^2) - (\mathbb{E}X_i)^2) = \frac{n-1}{n}\text{Var}\{X_1, \ldots, X_n\}.
\end{aligned}$$

□

2.2 Maximal likelihood estimators

Definition 2.4: Let θ be the parameter for a distribution and $x^n = (x_1, \ldots, x_n)$ the sample. The **likelihood** of x^n given θ is

$$\mathcal{L}(\theta) := \mathbb{P}(x^n|\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

where $f(x;\theta)$ is the probability density of x given θ. The **log-likelihood** is

$$\ell(\theta) := \ln \mathbb{P}(x^n|\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta).$$

(We usually aren't too careful with constants, and may drop them.) The **maximal likelihood estimator** is

$$\arg_\theta \max \mathcal{L}(\theta) = \arg_\theta \max \ell(\theta).$$

2.2.1 Examples

We calculate the MLE for several distributions.

1. Bernoulli. Here each X_i is 0 or 1. We have

   $$\mathcal{L}(\theta) = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}, \qquad \ell(\theta) = (\sum X_i)\ln p + (n - \sum X_i)\ln(1-p).$$

   Setting $\frac{\partial^2 p}{\partial =^2} 0$ gives $\widehat{p} = \frac{\sum X_i}{n}$ as MLE.

2. Normal. Finding the MLE for μ means maximizing $\ell(\mu,\sigma) = C - \sum \frac{(x_i-\mu)^2}{2\sigma^2}$, which is minimizing the sum of squares

   $$\sum (x_i - \mu)^2.$$

   The MLE is $\widehat{\mu} = \frac{1}{n}\sum x_i$.[6]

3. (Uniform distribution) Given $\mathcal{F} = \{U_{[0,l]}\}$, the estimated l is $\min\{X_1, \ldots, X_n\}$. (Suppose the buses in an unknown city are labeled 1 to N. Assuming no knowledge of a prior distribution on number of buses, and you see buses numbered $b_1, \ldots, b_n$, your best guess for the number of buses is $\max\{b_1, \ldots, b_n\}$.)

4. Linear regression. We assume x_i are fixed, and $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ is error. Then we want to minimize

   $$\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

   (minimize least squares). Setting $\frac{\partial^2 \beta_0}{\partial =^2} \frac{\partial^2 \beta_1}{\partial =^2} 0$ gives the system $\begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i X_i \end{pmatrix}$ which has solution

   $$\beta_1 = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2} \qquad \beta_0 = \overline{Y} - \beta_1 \overline{X}.$$

[6] It does NOT make sense to find the MLE for σ, which is ∞.

5. Multivariate linear regression. Here $Y = X\beta + \varepsilon$ where X is $n \times p$. The MLE is again given by least squares, which is given by the projection $\widehat{\beta} = (X^T X)^{-1} X^T Y$ (assume X has full rank; this is necessary).

2.2.2 Properties of MLE

Define some quantities first.

Definition 2.5: 1. **KL distance**

$$D(f, g) = \int f(x) \ln \left(\frac{f}{g} \right) \, dx.$$

Why do we care about this? Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n} \sum_i \ln \frac{f(X_i; \theta)}{f(X_i; \theta_*)}$$

which has the nice property that the maximum is 0. (Without the $\frac{1}{n}$ it would blow up.) By LLN the expected value of this is exactly $-D(\theta_*, \theta)$.

2. **score function** $s(X; \theta) = \frac{\partial \ln f}{\partial \theta}$.
   Important property: $\mathbb{E}s = \int_{-\infty}^{\infty} s(X; \theta) f \, dx = (\int_{-\infty}^{\infty} f \, dx)_\theta = 0$.
3. **Fisher information** $I(\theta) = \mathrm{Var}_\theta(s(X; \theta))$, $I_n(\theta) = nI(\theta)$. I.e., $I(\theta) = -\mathbb{E}((\ln f)_{\theta\theta})$.

1.

Theorem 2.6 (Convergence of MLE)**:** Suppose

(a) $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$,
(b) for all $\varepsilon > 0$, $\sup_{|\theta - \theta_*| \geq \varepsilon} M(\theta) < M(\theta_*)$.

Then the MLE $\widehat{\theta_n} \xrightarrow{P} \theta_*$.

Proof. First show that $M(\theta_*) - M(\widehat{\theta_n}) \xrightarrow{P} 0$. Then use continuity of M. □

2.

Theorem 2.7 (Asymptotic normality of MLE)**:** (a) $\mathrm{se} \sim \sqrt{\frac{1}{nI(\theta)}}$ and $\frac{\widehat{\theta_n} - \theta}{\mathrm{se}} \to N(0, 1)$.

(b) $\widehat{\mathrm{se}} = \sqrt{\frac{1}{nI(\widehat{\theta_n})}}$: why are we redefining $\widehat{\mathrm{se}}$? We defined it a different way before. Do these definitions coincide? $\frac{\widehat{\theta_n} - \theta}{\widehat{\mathrm{se}}} \to N(0, 1)$.

Proof. (a) Linearize to find that

$$\ell'(\widehat{\theta}) - \ell'(\theta) \approx (\widehat{\theta} - \theta)(\ell''(\theta)) \implies -\frac{\ell'}{\ell''}(\theta) \approx \widehat{\theta} - \theta.$$

Now

$$\sqrt{n}(\widehat{\theta_n} - \theta) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \to \frac{N(0, I(\theta))}{I(\theta)} \to N(0, 1),$$

the top in distribution, the bottom in probability. (The top uses CLT on $\sum(\ln f)_\theta$; the bottom uses LoLN on $\sum -(\ln f)_{\theta\theta}$.)

(b) Show that $\sqrt{\frac{I(\widehat{\theta_n})}{I(\theta)}} \xrightarrow{P} 1$.

□

3. Think of this as a chain rule.

   Theorem 2.8: If $\tau = g(\theta)$ and $g'(\theta) \neq 0$, then $\frac{\widehat{\tau_n} - \tau}{\widehat{\text{se}}(\widehat{\tau})} \to N(0, 1)$ where $\widehat{\tau_n} = g(\widehat{\theta_n}), \widehat{\text{se}}(\widehat{\tau_n}) = |g'(\widehat{\tau})|\widehat{\text{se}}(\widehat{\tau_n})$.

   Proof: just expand g using g'.

4. (Equivariance) If $\tau = g(\theta)$ is 1-to-1, then $\widehat{\tau_n} = g(\widehat{\theta_n})$. Follow definitions!

Write $x^n \leftrightarrow y^n$ if $f(x^n; \theta) = cf(y^n; \theta)$ as functions of θ. T is sufficient if $T(x^n) = T(y^n) \implies x^n \leftrightarrow y^n$. T is minimally sufficient if it is also a function of every other sufficient statistic. Factorization gives that $f(x^n; \theta) = g(t(x^n); \theta)h(x^n)$.

Some stuff on exponential families at end of §9.

3 Hypothesis testing

We are given hypothesis $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1 = \Theta_0^c$ and want to know which is true. (For example, $H_0 = \{\theta_0\}$ and $H_1 = \mathbb{R}\backslash\{\theta_0\}$.) H_0 is the **null hypothesis**; we reject or fail to reject it. In other words, we have a decision function δ that given X^n gives 0 or 1. There are two kinds of errors:

1. reject ($\delta = 1$) when H_0 is true
2. fail to reject ($\delta = 0$) when H_1 is true.

Definition 3.1: The **power** is

$$\beta(\theta) = \mathbb{P}(\delta = 1|\theta),$$

the probability of rejecting given θ. We want to maximize power for $\theta \in H_1$ (i.e., minimize $1 - \beta(\theta)$, which is the type 2 error).

The **size** is

$$\sup_{\theta \in H_0} \beta(\theta) = \sup_{\theta \in H_0} \mathbb{P}(\delta = 1|\theta).$$

The is the maximum probability of a false rejection (type 1 error). For example, when H_0 is $\{\theta_0\}$ this is just $\mathbb{P}(\delta = 1|\theta_0)$. We say a test has level α if it has size $\geq \alpha$.

Given rejection regions R_α of size α, the ***p*-value** is $\inf\{\alpha : T(x^n) \in R_\alpha\} = \sup_{\theta\in\Theta_0}(\mathbb{P}_{X^n}(T(X^n) \geq T(x^n)|\theta))$ where T is the statistic used in the test.

3.1 Basic tests

Exercise 3.2: 1. Give a hypothesis test for the normal distribution (z and t test).

2. Give a hypothesis test for comparing 2 means of normal variables when
   (a) 2 groups are independent
   (b) the elements are paired.

   Do the same for Bernoulli.

3. Compare the variances of 2 normal distributions.

Give a test comparing 2 variances of normal distributions.

Let z_α be the value of z such that $\int_z^\infty N(x)\,dx = \alpha$ where N is the standard normal.

For the examples below we take H_0 consisting of a single point, so the tests are double-tailed. For $H_0 = \{\theta \leq \theta_0\}$ and $\{\theta \geq \theta_0\}$ we use single-tailed tests.

1. For the normal distribution:
   (a) We know that $\lim_{n\to\infty} \frac{\widehat{\mu_n}-\mu_0}{\widehat{\text{se}}} \to N(0,1)$. Thus we can use the z-test which of size α:[7]
   $$\sqrt{n}\left|\frac{\widehat{\mu}-\mu_0}{\widehat{\sigma}}\right| > z_{\alpha/2}.$$
   The power is approximately $1 - \Phi(\frac{\mu_0-\mu}{\widehat{\sigma}} + z_{\alpha/2}) + \Phi(\frac{\mu_0-\mu}{\widehat{\sigma}} - z_{\alpha/2})$.
   (b) That is only an approximation, good when n large. The real distribution is the t-distribution with $n-1$ degrees of freedom.
   Now we use:

   Lemma 3.3: Let $X_1, \ldots, X_n \sim N(0,1)$. Then
   $$\sqrt{n-1}\frac{\widetilde{\mu}}{\widehat{\sigma}} \sim t_{n-1}.$$

   Proof. $\sqrt{n}\widehat{\mu} = (\frac{1}{\sqrt{n}} \ldots \frac{1}{\sqrt{n}})x$. Let A be an orthogonal matrix with first row $A_1 = (\frac{1}{\sqrt{n}} \ldots \frac{1}{\sqrt{n}})$. Let $y = Ax$. We have $y^T y = x^T x$ so
   $$n\widehat{\sigma} = x_1^2 + \cdots + x_n^2 - \underbrace{y_1^2}_{n\overline{X}^2} = y_2^2 + \cdots + y_n^2.$$

[7] It really doesn't matter for these approximate tests whether we use $\widehat{\sigma}$ or $\widehat{\text{se}} = \sqrt{\frac{n-1}{n}}\widetilde{\sigma}$ because $\sqrt{n-1} \sim \sqrt{n}$. For the exact tests we have to make the distinction.

This is distributed as χ^2_{n-1} and independent of y_1. Thus the distribution is $\sqrt{n-1}\frac{N(0,1)}{\chi^2_{n-1}} = t_{n-1}$. □

In summary we have

$$\sqrt{n-1}\frac{\widehat{\mu}-\mu_0}{\widehat{\sigma}} = \sqrt{n-1}\cdot \underbrace{\frac{\sqrt{n}(\widehat{\mu}-\mu_0)}{\sigma}}_{\sim N(0,1)} / \underbrace{\frac{\sqrt{n}\widehat{\sigma}}{\sigma}}_{\sim \chi^2_{n-1}} \sim t_{n-1}.$$

The more accurate test is

$$\sqrt{n-1}\left|\frac{\widehat{\mu}-\mu_0}{\widehat{\sigma}}\right| > t_{n-1,\alpha/2}.$$

2. Let $\mu_x = \mathbb{E}^m_{i=1} X_i$ and similarly for y.

   (a) First, the normal approximation. We have $\frac{\widehat{\mu_x}-\mu_x}{\widehat{\sigma_x}} \approx N(0,1)$ and similarly for y; clear denominators, add, and divide by standard deviation (noting variances add) to get

   $$\frac{\widehat{\mu_x}+\widehat{\mu_y}-\mu_x-\mu_y}{\sqrt{n\widehat{\sigma_x}^2+m\widehat{\sigma_y}^2}} \approx N(0,1);$$

   do the z-test.

   For an exact test, we find

   $$\frac{\widehat{\mu_x}+\widehat{\mu_y}-\mu_x-\mu_y}{\sqrt{n\widehat{\sigma_x}^2+m\widehat{\sigma_y}^2}} = \frac{N(0,\frac{\sigma_x^2}{m})+N(0,\frac{\sigma_y^2}{n})}{\sqrt{\sigma_x^2\chi^2_{m-1}+\sigma_y^2\chi^2_{n-1}}}$$

   which simplifies only when $\sigma_x = \sigma_y$ (assumption of equal variances) to get

   $$\sqrt{\frac{m+n}{mn}}\frac{N(0,1)}{\sqrt{\chi^2_{m+n-2}}}.$$

   Thus test

   $$\left|\sqrt{\frac{mn(m+n-2)}{m+n}}\frac{\widehat{\mu_x}+\widehat{\mu_y}}{\sqrt{n\widehat{\sigma_x}^2+m\widehat{\sigma_y}^2}}\right| > t_{m+n-2,\alpha}$$

   For unequal variances, use the Satterthwaite approximation (see 18.443 lecture 7).

   (b) Use the t-test on the pairs $x_i - y_i$.

   For Bernoulli, let $\widehat{\sigma} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}}$ and use the normal approximation to the binomial to test.

3. Suppose we have m, n samples respectively. Now if both variances equal σ^2,

   $$\frac{\widehat{\sigma_m}^2}{\widehat{\sigma_n}^2} \sim \frac{\frac{1}{m}\chi^2_{m-1}}{\frac{1}{n}\chi^2_{n-1}}.$$

   So take

   $$\frac{m(n-1)}{n(m-1)}\frac{\widehat{\sigma_m}^2}{\widehat{\sigma_n}^2}$$

   as the F-test statistic.

3.2 χ^2 goodness of fit

1. How to test multinomial distributions? Use the following

$$\sum_{j=1}^{r} \frac{(X_j - E_j)^2}{E_j}$$

as a test statistic for χ^2_{r-1}, where $E_j = np_j$ is the expected number in category j.

Proof. The covariance matrix A for the variables $\left(\frac{X_i - np_i}{\sqrt{np_i}}\right)_i$ has diagonal $(1-p_i)_i$ and off-diagonal entries $-\sqrt{p_i p_j}$. Note that $(\sqrt{p_i})_i$ spans the kernel. Note $A - I$ has rank 1 so all other eigenvectors have value 1. Let U diagonalize A; let $y = Ux$. Then

$$\sum_j \frac{(X_j - E_j)^2}{E_j} = x^T x = y^T y.$$

Because y has covariance matrix $\begin{pmatrix} 0 & 0 \\ 0 & I_{r-1} \end{pmatrix}$, this is χ^2_{r-1}. □

2. For goodness-of-fit of a sample to a distribution f, split the domain of f into intervals (where $\int f = \frac{1}{n}$, say), and run Pearson's test on them.

   Careful: here $\widehat{p_j}$ should be the MLE for the grouped distribution, the maximum for $\arg_\theta \max \mathbb{P}(I_1|\theta)^{v_1} \cdots$, rather than the MLE for the distribution, then grouped. (See 18.443 L12.)

   The degrees of freedom should be $r - \dim(\Theta) - 1$.

3. Give tests for independence and homogeneity.

   These are the same!

   Suppose there are N_{ij} observations for (i, j). We want $\max \prod p_i^{N_{i\bullet}} \prod_j q_j^{N_{\bullet j}}$. Take logs; the gradient should be in the span of $(1, \ldots, 1, 0, \ldots, 0)$ and $(0, \ldots, 0, 1, \ldots, 1)$ (Lagrange multipliers) so all the $N_{i\bullet}/p_i$ are the same, and similarly for q. df= $ab - (a-1) - (b-1) - 1 = (a-1)(b-1)$.

3.3 Nonparametric testing

For any distribution, let $F_n(x)$ be the estimated distribution function after getting n samples. By LLN, for a particular x, $F_n(x) \to F(x)$: $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x)))$.

The basis of nonparametric tests using F_n is the following:

Theorem 3.4: $\sup |F_n - F|$ does not depend on F.

Proof. Let $y = F_n(x)$ so that $x = F_n^{-1}(y)$; putting things in terms of y makes this into the problem for the uniform distribution. (Warning: some finesse is required because F_n can have jumps.) □

The Kolmogorov-Smirnov test uses

$$\mathbb{P}(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\leq t)\to H(t)=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}.$$

There are variations; see L14.

3.4 Multiple testing

If we do N tests at level α, we can expect pN false rejections. We can test at level $\frac{\alpha}{N}$ but often this is too stringent. A better way in practice is to order p-values from smallest to largest. If the ith p-value falls below $\frac{i\alpha}{m}$ then reject all tests below that.

4 Bayesian statistics