prod(big(71):100) \ prod(big(1):30)
binomial(big(100),30)
latex: \binom{n}{k}

Probability Models——a model of a random experiment

Omega — The space of outcomes
Random Variable, a function from Omega to R

Distributions — a property of a random variable

Independent random variables

We say that random variables $X_1, ..., X_n$ are independent if for any subsets $A_1, \ldots, A_n$ of $\mathbb{R}$, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n}(X_i \in A_i)\right) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

This is just saying that we can always multiply probabilities for individual random variables to get joint probabilities.

Expectation
The expectation of a random variable is its probability-weighted average value.

Linearity of Expectation
E[aX+bY] = aE[X] + bE[Y]

Variance
Var(X) = E[(X-μ)^2], where μ=E[X]

Covariance
Cov(X,Y) = E[(X-μx)(Y-μy)]

Var(X,X)=Cov(X,X)

Correlation — variance-normalized covariance, it measures to what extent the first/third quadrants “win out” over the second and fourth in this picture.

We cannot say anything definitively from a mere sample. It could be that the underlying measure was actually quite positively correlated, yet by sheer luck the points ended up showing this negative slant.

Bayes’ Theorem

$$\mathbb{P}(A \mid E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)} \qquad \mathbb{P}(A \mid E) = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E)}$$

P(A) - prior, probability of A
P(A|E) - posterior, updated probability after gaining the knowledge that E occured

P(E|A) - likelihood, how likely E was a priori, given that the event A occurs

Posterior is proportional to likelihood times prior

Bayes theorem for a distribution
Some code:
Binomial(n,p)
p^a(1-p)^b : beta(a+1,b+1), (function)
Beta(a,b), (quantile 的时候用), (distribution)

Conditional Expectation
First find the joint density along each vertical line
The conditional expectation of y is the density-weighted average value of y. The way to calculate a weighted average of a function is to multiply that function by the weighting function and integrate:
Integrate(y* joint_density, (y, 0, 1) )

Tower Law — averaging on average fives the average

E[ E [ Y | X ] ] = E [Y] $\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T \mid I]] = \frac{\mathbb{E}[T \mid I=1] + \mathbb{E}[T \mid I=2] + \mathbb{E}[T \mid I=3]}{3}$

Common Distributions

Bernoulli Distribution
It's a random variable which can realize at most two values.
We encode these values as 0 and 1. We customarily use the variable p to denote the amount of probability mass at 1.
Mean: p; Variance: p(1-p)

Binomial Distribution
The sum of n independent random variables which are each Bernoulli with parameter p.
Mean: np; Variance: np(1-p)

Geometric Distribution
The index of the first one in a sequence of independent Bernoulli(p)'s; Monte Carlo simulation
Mean: 1/p

Poisson Distribution
Let np = λ
Binomial(n, λ/n) and take n -> infinity
Mean: λ ; variance: λ

Exponential Distribution
The geometric distribution is to the exponential distribution as the binomial distribution is to the Poisson distribution.

Normal distribution
A multivariable standard normal random vector is a vector of independent standard normal random variables.
Since the Gaussian density has the special property, the multivariate standard normal density function

is radially symmetric.
A multivariate normal random variable is any linear transformation of a multivariate standard normal vector.

The key piece of information we need is that the covariance matrix of a mean-zero random vector is $\mathbb{E}[XX']$.

With this in mind, we can de-mean $AZ + \mu$ by pulling off the $\mu$ term and then get

$$\mathbb{E}[(AZ)(AZ)'] = A\mathbb{E}[ZZ']A' = AA'.$$

So the covariance matrix $\Sigma$ and the linear transformation $A$ are related by $\Sigma = AA'$.

Given a desired covariance matrix $\Sigma$, we can always take its square root and let $A$ be that matrix.

This works since $\Sigma$ is symmetric, and therefore its square root is also symmetric, and so we'll have $AA' = AA = A^2 = \Sigma$, as desired.

Independent sums of random variables

Central limit theorem: a sum of many independent random variables with a common distribution is approximately normally distributed (with appropriate mean and variance)

Example:

📌 Suppose that the percentage of residents in favor of a particular policy is 64%. We sample individuals uniformly at random from the population.

- In terms of $n$, find an interval centered at 0.64 such that the proportion of residents polled who are in favor of the policy is in $I$ with probability about 95%.

So that's it: we should use the interval

$$\left[0.64 - 1.96\frac{0.48}{\sqrt{n}},\ 0.64 + 1.96\frac{0.48}{\sqrt{n}}\right]$$

The random variable of interest (the proportion of respondents in favor) has a mean 0.64, so we just need to work out its variance. Letting $X_k$ be the $\{0,1\}$-valued response variable for the $k$th person, we get

$$\text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{n\text{Var}(X_1)}{n^2} = \frac{0.64\,(1-0.64)}{n}$$

In that last step, we're using the formula $p(1-p)$ for the variance of a Bernoulli.

This is going to be the variance of our random variable, and the central limit theorem tells us it's approximately normally distributed with that variance.

Simulation techniques — inverse CDF trick and multivariate Gaussian sampling

Cumulative distribution function: given a probability measure on the real number line, the corresponding CDF F is the function which maps each value x∈R to the amount of probability mass in the interval (−∞,x]

Based on where the graph is steepest to tell where the mass is most concentrated on the CDF graph.

The inverse CDF trick
Steep portions of the CDF occupy a lot of vertical space in the graph, while shallow portions occupy less. So we could choose a Uniform(0,1) random variable and place it along the vertical axis, and move horizontally till we hit the graph. The resulting point on the x-axis at least has the property that it is more likely to be in intervals with a lot of probability mass and less likely to be in intervals with less.

Generalized inverse of F: the map from each horizontal line to the x-value where the line hits the graph of F

To find the CDF corresponding to this PDF, we integrate it to get $\int_0^x nt^{n-1}dt = x^n$. The inverse of this function is the $n$th root function, so the formula we're going for here is $\sqrt[n]{U}$.

Exercise:
Sample from a Gaussian 2-vector whose components have variance 2 and 1 respectively and correlation -0.99
First, figure out the covariance matrix
Diagonal entries are the covariances of each component with itself, in other words, the variance of each component, so 2 and 1
Corr(X,Y) = cov(X,Y)/(standard deviation(x),standard deviation(y))
Sigma = [2 -0.99*sqrt(2); -0.99*sqrt(2) 1]

Then find the matrix representing the appropriate transformation of a standard multivariate normal
A = sqrt(sigma)
Histogram2d([Tuple(A*rand(Normal(0,1),2)) for _ in 1:1_000_000], ratio = 1)

Introduction to Statistics
Code: countmap (出现次数)
using StatsBase
countmap([mysteryRV() for _ in 1:100_000])

Example: WHich common distribution does an unknown function sample from?
tally = sort(countmap([mysteryRV2() for _ in 1:100_000]))
sticks(collect(keys(tally)), collect(values(tally)))

The advantages of using the parametric approach
If the distribution rally is approximately one of the ones in your parametric family, we'll get better results because we're able to leverage more data in the service of estimating the parameters.

The advantages of using the non parametric approach
If the distribution is not close to one in your parametric family, then you'll be estimating the distribution in a way that's systematically biased in the direction of that parametric family. There's no way for your estimate to be particularly close even if you have an enormous amount of data.

Kernal Density Estimation

To do this, we define $D_\lambda(u) = \frac{1}{\lambda} D\left(\frac{u}{\lambda}\right)$. The inner factor of $1/\lambda$ scales the function horizontally, and the outer one scales it vertically so as to ensure it represents the same amount of mass as it did before scaling.

Then we define $K_\lambda$ in terms of $D_\lambda$ as before, just multiplying copies of $D_\lambda$ along each axis: $K_\lambda(x, y) = D_\lambda(x)D_\lambda(y)$.

Conditional expectation function x->E[Y| X=x], call r(x), is a function which is entirely determined by the probability distribution of (X,Y)

Regression: function as a map from a set of probability measures to a set of functions
In short, probability measure in, function out.

Cross-validation — to find the best λ
Look at the average value obtained by leaving out a point and checking the resulting density at the left-out point. We can then choose the λ value which minimizes this average.
This will cause problem: this approach favors smaller λ values, because it can get a big boost for those values of λ from a handful of points which happen to be really nearby.

L^2 distance: One simple to fix it is to integrate the squared difference between the two functions.

Stone theory

**Theorem** (Stone)
Suppose that $f$ is a bounded probability density function. Let $\hat{f}_n^{\mathrm{CV}}$ be the kernel density estimator with bandwidth $\lambda$ obtained by cross-validation, and let $\hat{f}_n$ be the kernel density estimator with optimal bandwidth $\lambda_n^{\min}$. Then

$$\frac{\int (f-\hat{f}_n^{\mathrm{CV}})^2}{\int (f-\hat{f}_n)^2}$$

converges in probability to 1 as $n \to \infty$. Furthermore, there are constants $C_1$ and $C_2$ such that $\int (f-\hat{f}_n)^2 \approx C_1 n^{-4/5}$ and $\lambda_n^{\min} \approx C_2 n^{-1/5}$ for large $n$.

Nadaraya-Watson: Regression estimator, a local, kernel-weighted average of the nearby observations' y-values

When λ is large, we're taking an weighted average where all of the weights are approximately the same. In other words, just a straight average.
When λ is small, there will be vertical lines which don't intersect the mass at all, and we get breaks in the graph (Where NaNs are being returned as a result of division by zero)

Point estimation — the single number we're trying to estimate is a single point on the number line

Statistical functional: a function from the space of distributions to the real number line

An estimator is any random variable which is a function of n i.d.d draws from a distribution v. It needs to be specified explicitly which statistical functional T the estimator is intended as an estimator for.

The plug-in estimator
Empirical distribution: puts mass 1/n at each of the n observations
Plug the empirical measure v_hat into the statistical functional T.

The empirical range is always smaller than the actual range for a random variable whose distribution has a density
Because: there's always going to be a small gap at the top and at the bottom between the actual min and the least observation and between the actual max and the largest observation.

Bias
An estimator is biased if its bias is nonzero and unbiased if its bias is zero

More precisely, the **bias** of an estimator $\hat{\theta}$ of a statistical functional whose actual value is $\theta = T(\nu)$ is defined to be

$$\mathbb{E}[\hat{\theta}] - \theta$$

Standard error: the standard deviation of an estimator

Mean squared error

**Definition** (Mean squared error)
The mean squared error of an estimator $\hat{\theta}$ is $\mathbb{E}[(\hat{\theta} - \theta)^2]$.

Intuitively the MSE of an estimator is large if its bias is large or if its variance is large.

More precisely, the mean squared error is exactly the sum of the squared bias and the variance of the estimator:

$$\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{variance}$$

Consistent

**Definition** (Consistent)
An estimator is **consistent** if $\hat{\theta}$ converges to $\theta$ in probability as $n \to \infty$.

Bessel's correction, multiply by n/(n-1) to get an unbiased estimator (Sample variance)

Confidence intervals

**Definition** (Confidence interval)
Consider an unknown probability distribution $\nu$ from which we get $n$ independent observations $X_1, \ldots, X_n$, and suppose that $\theta$ is the value of some statistical functional of $\nu$.

A **confidence interval** for $\theta$ is an interval-valued function of the sample data $X_1, \ldots, X_n$. A confidence interval has **confidence level** $1 - \alpha$ if it contains $\theta$ with probability at least $1 - \alpha$.

MLE
We need to aggregate these values somehow to get a single score which tells us how good the fit is. Since the observations are unde3rstood to be drawn from the distribution independently of one another, we multiply these individual likelihood values to get the overall likelihood.

Short coming: We can get under flow if we multiply together too many small values. For example, if we multiply 1/2 by itself just a couple thousand times, we get down well below 2^(-1074) which is the smallest positive representable number in the Float64 system.

Fix it by log.
1. Taking the log of a product is the same as summing the logs of each factor
2. Maximizing the logarithm of an expression is the same as maximizing the expression itself, that's because log is a monotone function

In conclusion, maximizing the log likelihood to determine the best fit, the sum of the logs of the density values evaluated at each observation

MLE is biased, consistent, both the bias and the variance converge to zero as the number of observations go to infinity
Approximately normally distributed with mean theta, helpful for calculating confidence intervals.
Asymptotically optimal, you aren't going to find another estimator which converges faster to the

correct value theta than the MLE, as the number of observations goes to infinity.

The MLE drawbacks

1. Computational obstacles. Nothing guarantees tractability of the MLE optimization problem. We;re going to see this much later in the course where we find a sum of multiples of Gaussian densities to fit a multimodal distribution (Like the one we saw earlier today which we couldn't fit with a single Gaussian).
2. Misspecificaiton. We saw this one earlier too: if your distribution isn't actually in your parametric family, your estimator might be unreasonable.
3. Unbounded likelihood. It is possible in some cases to get arbitrarily large likelihood values.
4. It doesn't account for prior information.

Hypothesis Test

Statistically significant — How big of a difference is

1. We state a hypothesis H0 — null hypothesis
2. Come up with a test statistic T, a function of the data X1, ……, Xn, and for which we can evaluate the distribution of T assuming the null hypothesis.
3. Alternative hypothesis Ha, under which T is expected to be significantly different from its value under H0
4. Significance level (like 5% or 1%), and based on Ha we determine a set of values for T — called the critical region — which T would be in with probability at most α under the null hypothesis.
5. After determining all of the above, we run the experiment, evaluate T on the samples we get, and record the result as t_obs
6. If t_obs falls in the critical region, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

The corresponding p-value is defined to be the minimum α-value which would have resulted in rejecting the null hypothesis, with the critical region chosen in the same way.

T-test — one type of hypothesis test