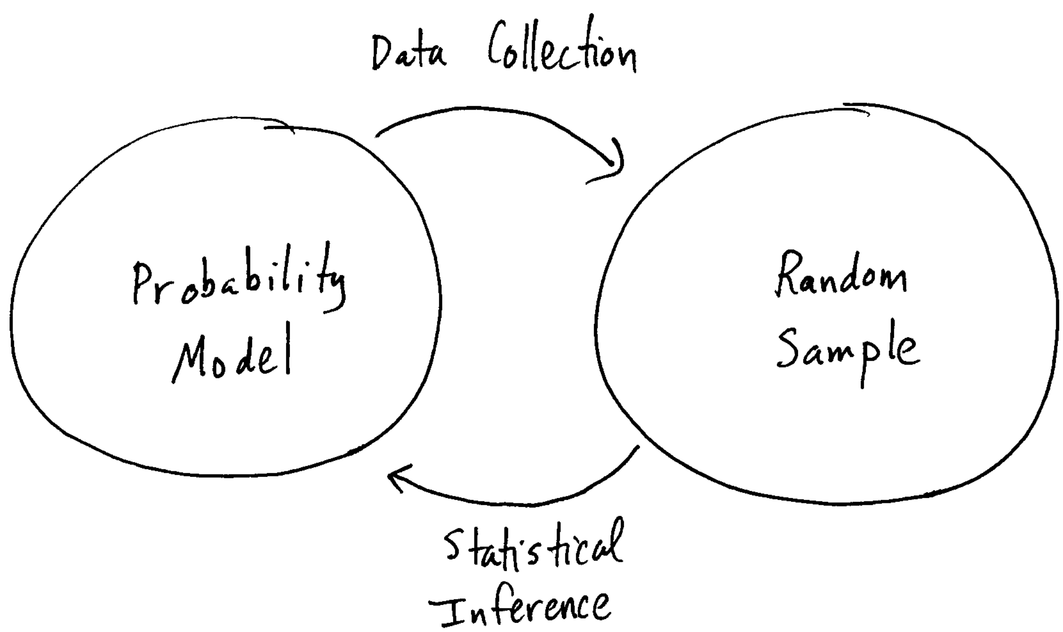
STAT 310 REVIEW

I. Basic Probability

II. Random Variables and Distributions

III. Statistical Inference (Estimation / Testing)

The Big Picture



Probability model = family of distributions parameterized by a common parameter (e.g., normal, Poisson, ...)

I. Basic Probability

At the heart of probability and statistics is the notion of an <u>experiment</u>, which is a (repeatable) phenomenon whose result is not known in advance.

Key concepts:

- <u>outcome</u>: result of an experiment
- <u>event</u>: collection of outcomes
- <u>outcome/sample space</u>: set of all possible outcomes.

A <u>probability law</u> P is a function that assigns events to probabilities (real numbers between 0 and 1). and satisfies the <u>axioms of probability</u>

Equally Likely Outcomes

Probability law that assigns each outcome the same probability. (Only possible when outcome space is finite.) In this case, computing the probability of an event boils down to <u>counting</u>:

$$P(A) = \frac{N(A)}{N(S)}.$$

If an experiment is ~~comprised of~~ conducted by repeating a simpler experiment several times, the manner of counting hinges on whether the sampling occurs <u>with</u>/<u>without</u> <u>replacement</u> and whether the outcomes (of the simpler experiment) are viewed as <u>ordered</u> or <u>unordered</u>.

Other important concepts :

- conditional probability
- independent events
- Bayes theorem

II. $\boxed{\text{Random Variables and their Distributions}}$

A random variable is a function that maps outcomes of an experiment to real numbers.

$\boxed{\text{Key point}}$: Sometime the outcome of an experiment is already a real number, in which case we view that outcome as a random variable.

At other times, the outcomes of an experiment are non-numeric and a "measurement" of the outcome is the random variable. (e.g., weather pattern $\longrightarrow$ temperature).

For our purposes, however, we can usually disregard the underlying experiment and view a random variable as a function X that assumes ~~certain~~ different values with different probabilities.

Two main kinds of RVs, and examples

1. Discrete :
   Bernoulli, binomial, Poisson, neg. binomial, hypergeometric, geometric, discrete uniform

2. Continuous :
   Uniform, exponential, normal, χ^2, gamma, t, Laplacian

Key concepts : pmf / pdf, cdf, MGF, mean, variance, moments, expectation, (lower) percentiles, upper percentiles, distribution function technique, change of variable technique.

A function of a random variable is also a random variable.

If $Y = \varphi(X)$, and

- the pdf of X is $f(x)$
- the pdf of Y is $g(y)$

then $E[Y]$ can be computed in 2 ways:

- $E[Y] = \int \varphi(x) f(x)\, dx$
- $E[Y] = \int y\, g(y)\, dy$

The two are equivalent. We have assumed this throughout the course, but it is actually non-trivial to prove.

Extensions:

- bivariate / multivariate distributions
- joint / marginal pmf / pdf
- covariance / correlation
- independence

III. Statistical Inference

In statistical inference, the true distribution governing a random variable is unknown. The goal is to make an inference about the true distribution by collecting a random sample.

Two major kinds of inference:

- estimation
- testing

Before we can make any headway, we must first make an assumption about the family of distributions from which the data arise. This is our probability model.

The probability model is usually based on knowledge of the physical process generating the data. For example, if X is the number of times we need to flip a coin until 5 heads are observed, X is negative binomial. (even if the prob. of heads is unknown).

Once a probability model is specified, the next step is to devise a statistic on which to base the inference, and then determing its (sampling) distribution.

If we cannot determine the sampling distribution, then our procedure is somewhat arbitrary, i.e., we have no way of assessing the performance of the inference procedure. (For example, we need to know the sampling distribution to assign confidences to interval estimates or levels of significance to tests.)

Fortunately, when the data is normal, we can determine the sampling distributions of (and related to) $\bar{X}$ and S^2. Here the relevant distributions are normal, χ^2, and t. Furthermore, by the central limit theorem, many of our results ~~[illegible]~~ hold approximately even when the sample is not normal.

Estimation

A. Point estimation

Maximum likelihood : ~~solve~~ maximize

$L(\theta) = L(\theta; x_1, \dots, x_n)$ with respect

to θ.

Unbiased : $E[\hat{\theta}] = \theta$.

B. Interval Estimation

$$P(\theta \in [a, b]) = 1 - \alpha,$$

where a, b depend on $x_1, \dots, x_n$ and α

and any known parameters.

confidence $= 100(1-\alpha)\%$

__Testing__ Let $\mathcal{P}$ be a property of the distribution governing the data.

$$H_0 : \mathcal{P} \text{ is true}$$

vs.

$$H_1 : \mathcal{P} \text{ is not true}$$

The goal is usually to devise a test whose __level of significance__ is known.

* Sometimes only H_0 is specified. Rather than specify an alternative, it is often customary to decide between these two options:

  - reject H_0
  - fail to reject H_0

Tests are usually named after the distribution of the test statistic. Examples: t-test, χ^2-test. There are several kinds of t-tests and χ^2-tests.

Connection between Confidence Intervals and Tests

Recall that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known, then

$$\left[\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ confidence interval. Meanwhile, when testing $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$, the test with level of significance α accepts H_0 when

$$\bar{X} \in \left[\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}},\ \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right].$$

Notice that these results parallel each other.

This is not a coincidence.

A similar correspondence holds when σ^2 is unknown or when we are interested in one-sided confidence intervals / tests.

Key concepts

- MLE
- unbiased point estimator
- confidence intervals
- hypotheses (simple/composite, one-sided/two-sided)
- Type I/II errors
- p-value
- CLT
- distributions of $\bar{X}$, S^2, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, $\frac{\bar{X}-\mu}{s/\sqrt{n}}$

  when data are normal

- t-tests, χ^2 tests.
- sums of independent RVs.

In addition to the material we have covered in lecture / homework, I think it would be a very good way to study if you read the following sections:

6.6
7.3, 7.4, 7.5
8.3

Important sections we didn't have time to cover:

- regression (7.8, 7.9)
- analysis of variance (8.6, 8.7)

These will not be on the exam, but if you're eager to learn more statistics, start here.