Probability and Statistics Review Continued…

Some Distributions

Discrete Distributions

- A discrete random variable,X, takes on only a finite (or countably infinite) number of values
- Defined by probability mass function (PMF),p

$$p(x_i) = P(X = x_i) \text{ and } \sum_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} = 1$$

Assigning Probabilities: Counting

For a discrete, finite sample space

$$P(A) = \frac{\#\,\text{ways to get event A}}{\#\,\text{possible events}}$$

Example: after throwing two fair dice, what is the probability that the outcomes sum to 7?

$$P(7) = \frac{6}{6^2} = \frac{1}{6}$$

Counting Rules

- The Multiplication Rule: If an event has M independent steps, each step I has n_i possibilities, then the total number of possibilities is n_1*n_2*n_3*…*$n_{M.}$

- Permutations: The number of permutations (<u>ordered</u> samples) of k objects selected from N distinct objects is ("sampling without replacement"):

$$P_k^N = \frac{N!}{(N-k)!} = N*(N-1)*(N-2)*...*(N-k+1)$$

- Combinations: the number of ways an <u>unordered</u> subset (k) of objects can be selected from N objects ("N choose k"):

$$\binom{N}{k} = \frac{P_k^N}{k!} = \frac{N!}{(N-k)!k!}$$

Bernoulli Random Variable

- A Bernoulli random variable takes on only two values: 0 or 1

$$p(1) = p$$

$$p(0) = 1 - p$$

$$p(x) = 0, \text{ if } \mathrm{x} \neq 0 \text{ and } \mathrm{x} \neq 1$$

- A Bernoulli trial is like flipping coin

Binomial Random Variable

- Suppose we perform n independent bernoulli trials with $p(1)=p$
- The probability of K 1s ("heads") is a binomial random variable:

Binomial Random Variable

- Suppose we perform n independent bernoulli trials with p(1)=p
- The probability of K 1s ("heads") is a binomial random variable:

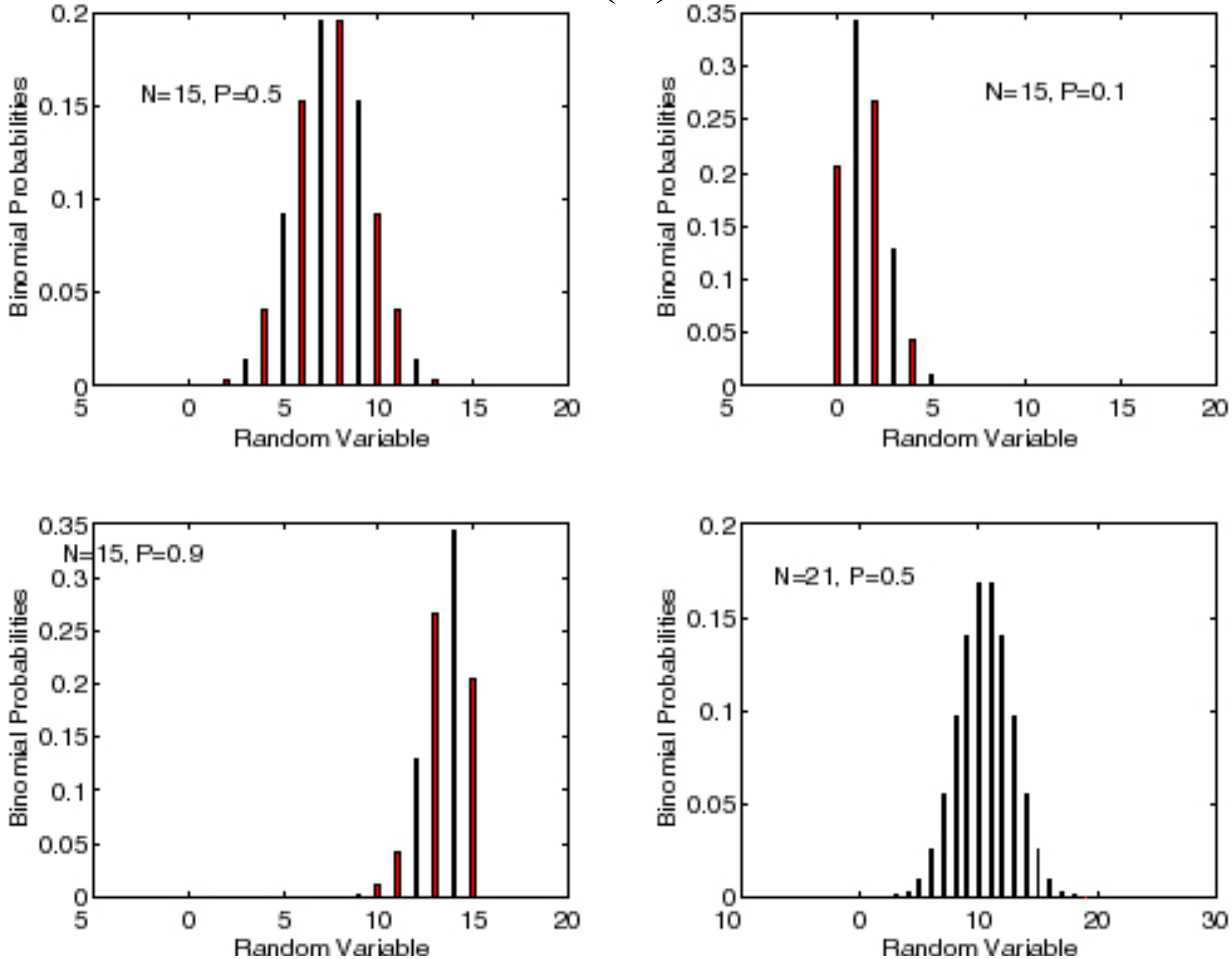
$$P(k\ \text{heads}) = \binom{\mathrm{N}}{\mathrm{k}} p^{k}(1-p)^{N-k}$$

$$\mu = Np$$

$$\sigma^{2} = Np(1-p)$$

Binomial PDF

$$P(k \text{ heads}) = \binom{N}{k} p^k (1-p)^{N-k}$$



Multinomial Distribution

- The multinomial distribution generalizes the binomial to more than two outcomes
- Imagine n experiments, where each experiment has k possible outcomes each with p_i
  - Example: rolling a fair die, $p_i = 1/6$
- The probability of counts $X=\{x_1, x_2, \ldots x_k\}$ is:

$$P(X) = \begin{cases} \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} & \text{when } \sum_{\mathrm{i=1}}^{\mathrm{n}} x_i = n \\ 0 & \text{otherwise} \end{cases}$$

The Poisson Distribution

If an event happens with a rate of λ events in some interval

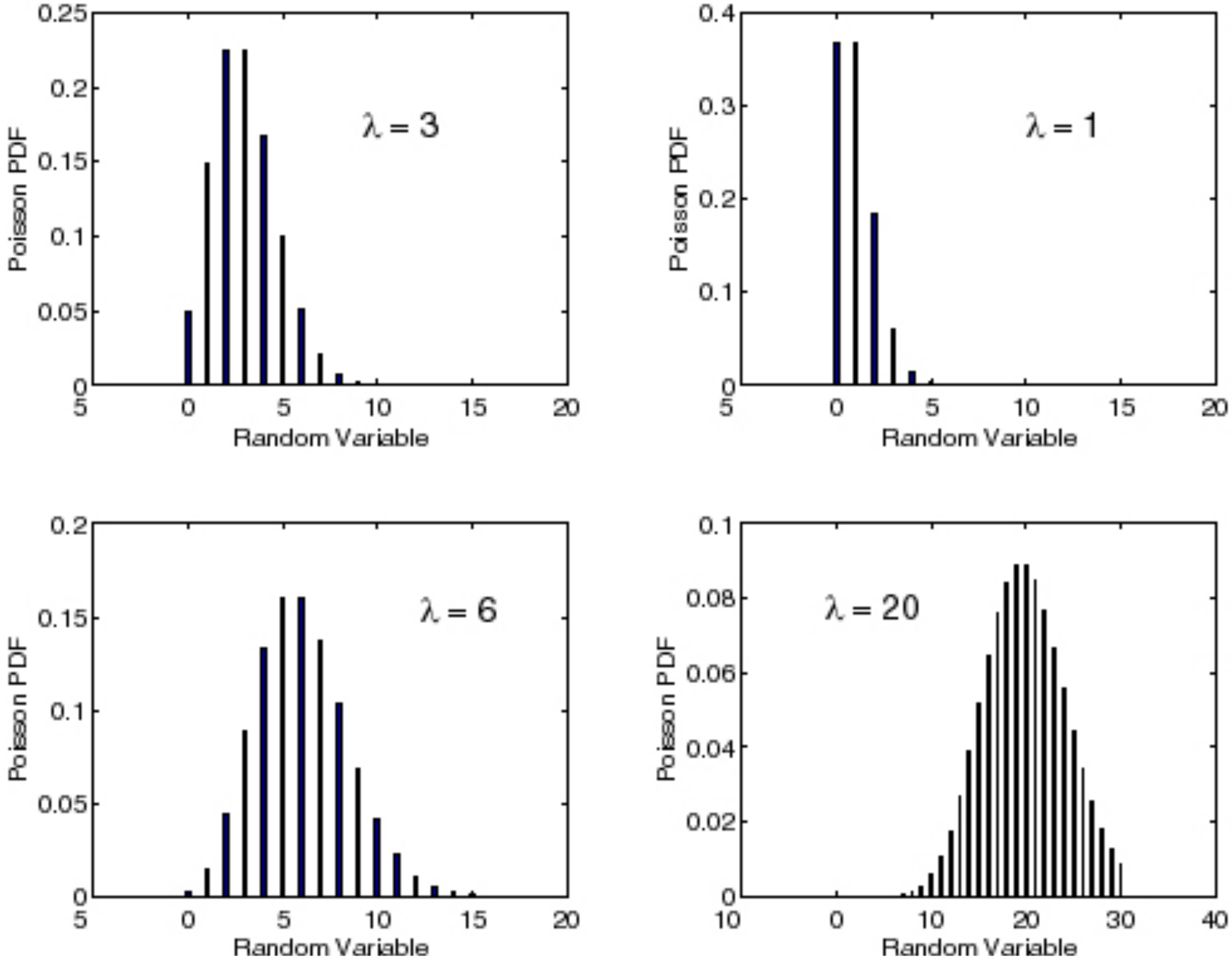
The probability of its occurring k times in the interval is:

$$P(K) = \frac{e^{-\lambda}\lambda^{k}}{k!} \quad k = 0,1,2...$$

$$u = \sigma^2 = \lambda$$

The Poisson PDF

$$P(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$



Poisson Example

A short DNA probe (10-mer) has a probability of 0.001 of hybridizing with a DNA 10-mer.

Now assume that the nucleotide composition of a 1-kb-long genomic DNA enables the number of binding sites of this probe to be modeled with a Poisson distribution.

What is the probability of having two or more sites for this probe in this genomic DNA?

Useful Discrete Distributions

- Binomial
- Multinomial
- Poisson
- Geometric
- Hypergeometric
- Negative Binomial

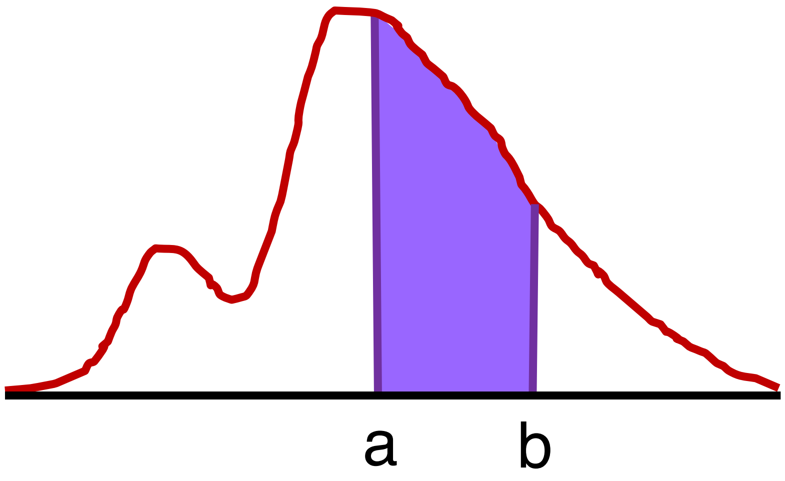
Continuous Distributions

- Continuous random variables take on continuum of values
- Characterized by a probability density function (PDF), f(x):

$$f(x) \geq 0$$

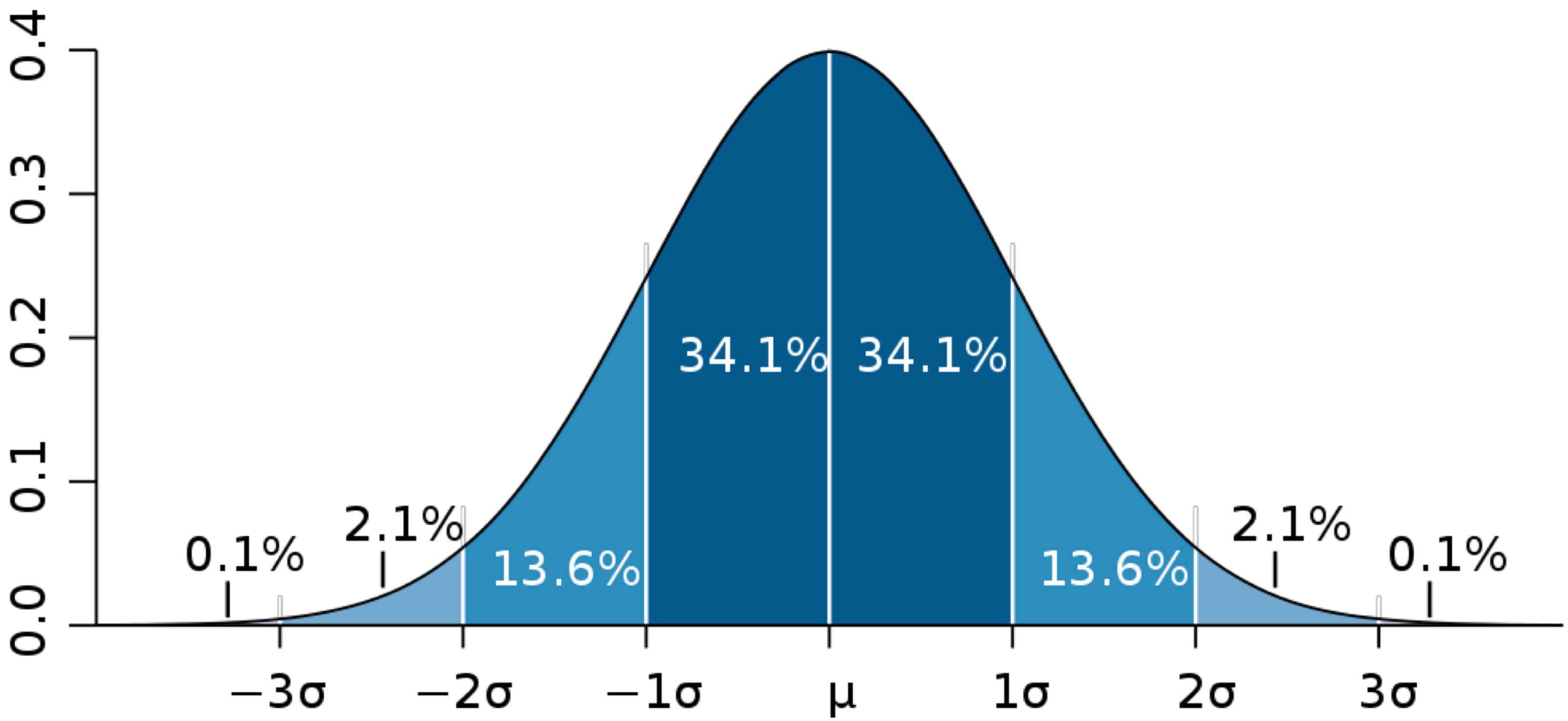
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$



Gaussian (Normal) Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Properties of Normal Distribution

- Linearity
  - If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, then $Z = X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$ If
  - $X \sim N(\mu, \sigma^2)$ then $Z = aX + b \sim N(a\mu+b, a^2\sigma^2)$
  - If $\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathrm{Z} = \dfrac{\mathrm{X} - \mu}{\sigma} \sim N(0,1)$
  - Z is called a standard normal variable
- Maximum entropy distribution (later lecture)
- Uncorrelated => Independent (if jointly normal)
  - If P(A,B) ~N, and A is uncorrelated with B, then $A \perp B$
- The normal distribution is **common**…

The Central Limit Theorem

The mean (times $\sqrt{n}$) of N statistically independent random variables has (under almost all circumstances when N is above about 10) a probability distribution that is well approximated by a Gaussian distribution function.

Central Limit Theorem in Action:

http://www.stat.sc.edu/~west/javahtml/CLT.html

http://www.rand.org/methodology/stat/applets/clt.html

χ^2 distribution

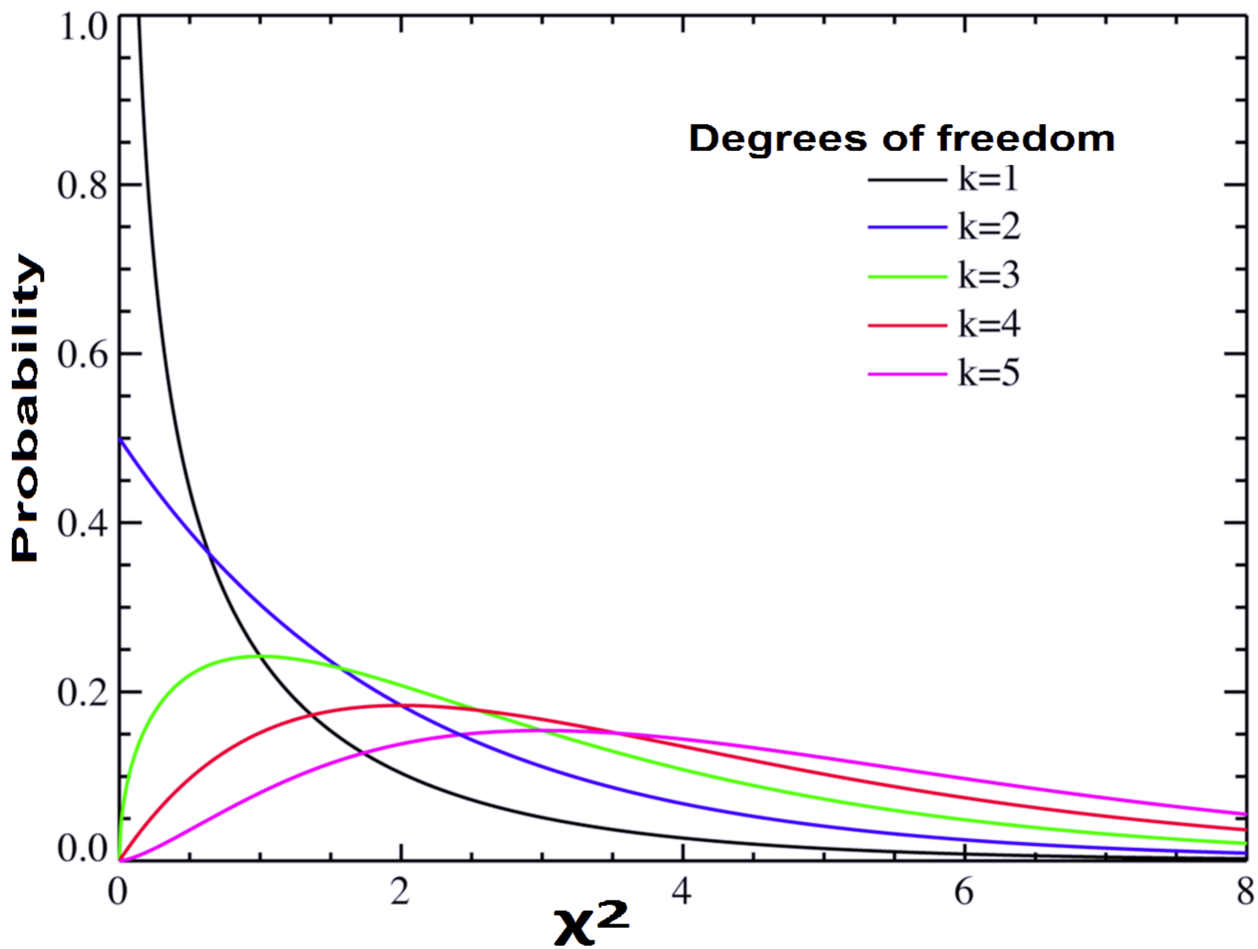
$Z_i \sim N(0,1)$ Z_i is standard normal

$$U = \sum_{i=1}^{n} Z_i^{\,2}$$

Follows a χ^2-distribution with n degrees of freedom (dof) *

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}$$

χ^2 distribution PDF



Student t distribution

$Z \sim N(0,1)$ Z is standard normal

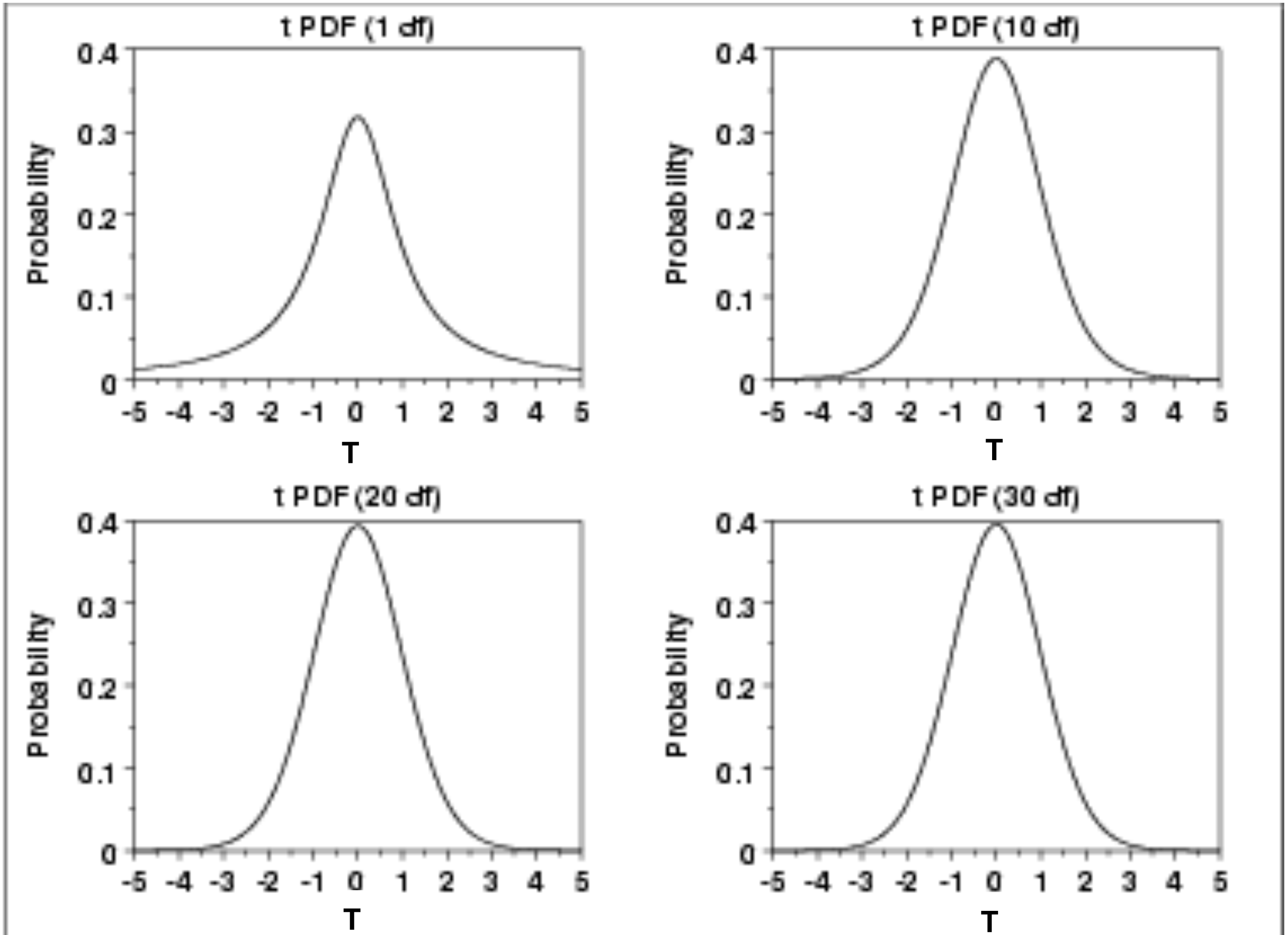
$U \sim \chi_n^2$ U is chi-squared, n dof

$Z \perp U$ Z and U are independent

$$\boxed{Z / \sqrt{U/n} \quad \textbf{Follows a t-distribution with n dof}}$$ *

$$f(t) = \frac{\Gamma\left[(n+1)/2\right]}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

The PDF of the t-Distribution



Useful Continuous Distributions

- Gaussian
- χ^2 distribution
- Student t distribution
- F distribution
- Log normal distribution
- Exponential distribution

- Conjugate Prior distributions
  - Beta (conjugate to Bernouli, binomial)
  - Gamma (Poisson, exponential)
  - Dirichlet (multinomial)

  => More in Classification Lecture

Expected Values

The expected value of a random variable is its probability weighted average value:

$$E[X] = \int_{-\infty}^{\infty} x p(x) dx$$

We can define the expected value of any function f(X) of X:

$$E[X] = \int_{-\infty}^{\infty} f(x) p(x) dx$$

One important expectation is the variance:

$$Var(X) = E\left[(X-\mu)^2\right] = E\left[X^2\right] - E[X]^2$$

Covariance and Correlation

- We can take the expected value of a function of two random variables:

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p(x,y)dxdy$$

- We can then define the covariance and correlation of X and Y as:

$$\text{Covariance}_{XY} = E[(X - E[X])(Y - E[Y])]$$

$$\text{Correlation}_{XY} = E[(X - E[X])(Y - E[Y])]/(\sigma_X \sigma_Y)$$

Probability and Statistics Review

What is Statistics?

Statistics is applied probability

- Statistics starts with data (samples)
- Generate a probability model or formulation from data
- Use probability calculus to make inferences about the data-generating process

Inferences

- Parameter Estimation
- Hypothesis Testing
- Many, many, others we will not cover…

Estimation

Parameter Estimation

Probability distributions, and more generally probabilistic models, typically depend on parameters, θ

Estimation is the problem of selecting parameters *consistent* with data

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

mean

variance

Forced Choice

Given the following samples from a normal distribution:

1, 0.8, 1.2, 1.1, 0.9, 0.7, 1.2

What would you choose if you had only the following two choices?

μ=1.1 μ=2.1

Maximum Likelihood Method

Suppose that random variables $X_1,\ldots,X_n$ have a distribution parameterized by θ:

$$P(x_1,\ldots,x_n | \theta)$$

The maximum likelihood approach selects θ according to:

$$\arg\max_{\theta} P(X_1,\ldots,X_n \mid \theta)$$

Note that here the X are given, and θ is unknown

Likelihood Function

Because X are no longer random we define a **likelihood function** (a function of θ):

$$L(\theta)=P(X_1,\ldots,X_n| \theta)$$

And we will find it convenient to maximize the **log likelihood**:

$$l(\theta) = \ln L(\theta)=\ln P(X_1,\ldots,X_n| \theta)$$

Maximum Likelihood Method

Suppose that random variables $X_1,\ldots,X_n$ have a distribution $P(x_1,\ldots,x_n | \theta)$,

Define the log likelihood function:

$l(\theta) = \ln L(\theta) = \ln P(X_1,\ldots,X_n | \theta)$

And choose θ such that

$$\frac{\partial l(\theta)}{\theta_i} = 0 \text{ for all i}$$

Example: Gaussian Distribution

If $X=X_1,\ldots,X_n$ are independent and $N(\mu,\sigma^2)$, then

$$f(X_i \mid \mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2}$$

$$\mathrm{l}(\mu,\sigma)=\ln f(X_i \mid \mu,\sigma)=-n\log\sigma-\frac{n}{2}\log 2\pi-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-u)^2$$

Setting partial derivatives to zero:

$$\frac{\partial l}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \qquad \boxed{\Rightarrow \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i}$$

Sample Mean

$$\frac{\partial l}{\partial\sigma}=-\frac{n}{\sigma}+\sigma^{-3}\sum_{i-1}^{n}(X_i-\mu)^2 \qquad \Rightarrow \hat{\sigma}^2=\frac{1}{n}\sum_{i-1}^{n}(X_i-\hat{\mu})^2$$

Properties of Estimators

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ **is an estimator for the parameter μ**

- Estimators are functions of random input samples
- Estimator are therefore random variables!
- Thus, estimators have expected values:

$$\text{For, } \overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathrm{E}\left[\overline{\mathrm{X}}\right] = \mu,$$

*

$$\mathrm{Var}\left[\overline{\mathrm{X}}\right] = \sigma^2 / n$$

If X_i are normal, then $\bar{X}$ is also normal

Properties of Estimators

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ **is an estimator for the parameter μ**

Estimators can have the following properties

- Consistency: A n->∞, the estimator converges to the correct value (in probability)
- Unbiased: E[estimator]=true value
- Efficient: low mean squared error of all estimators

The sample mean is consistent, unbiased, and efficient.

Sample Variance

The estimator for the sample variance derived above is biased:

$$E\left[\frac{1}{n}\sum_{i-1}^{n}\left(X_i-\hat{\mu}\right)^2\right]<\sigma^2$$

An unbiased estimator of variance is

$$S^2=\frac{1}{n-1}\sum_{i-1}^{n}\left(X_i-\hat{\mu}\right)^2$$ *

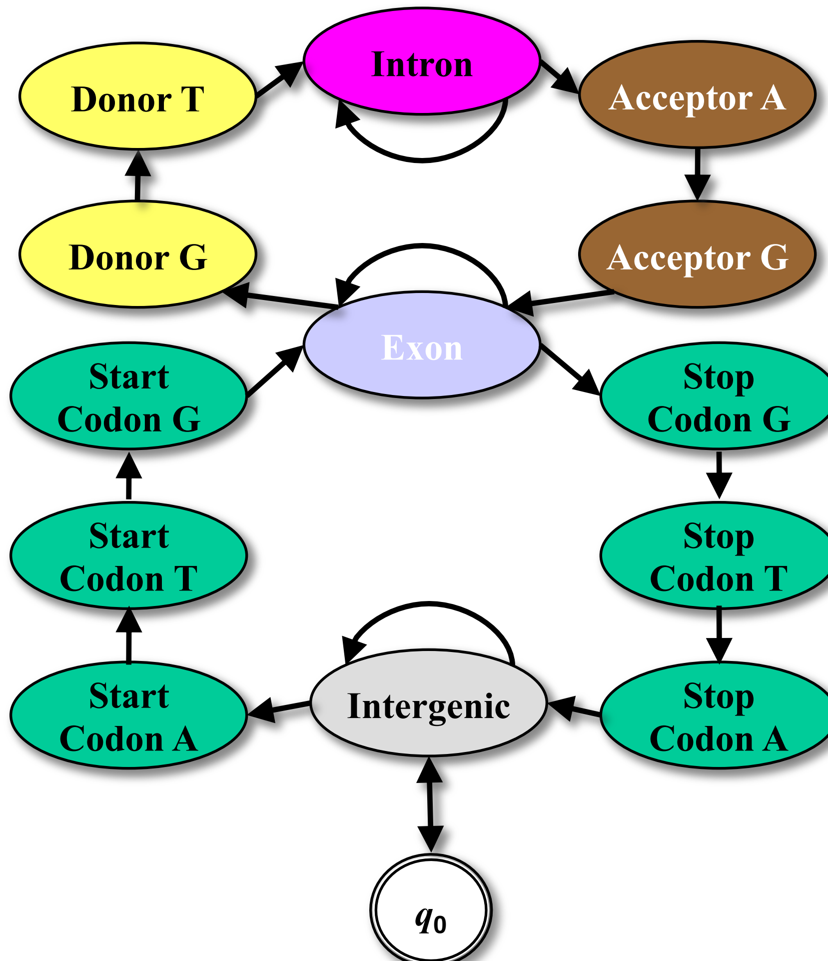
$$\left(n-1\right)S^2/\sigma^2\sim\chi^2_{n-1}$$ *

Proof at http://biology.ucf.edu/~pascencio/classes/Methods/Proof%20that%20Sample%20Variance%20is%20Unbiased.pdf

Remember Maximum Likelihood!

- It is the foundation for much of the modeling we will do in the course (e.g. HMMs)

- We will also extend this principle later using Bayes Rule (e.g. MAP estimators and classification)



Hypothesis Testing

Example

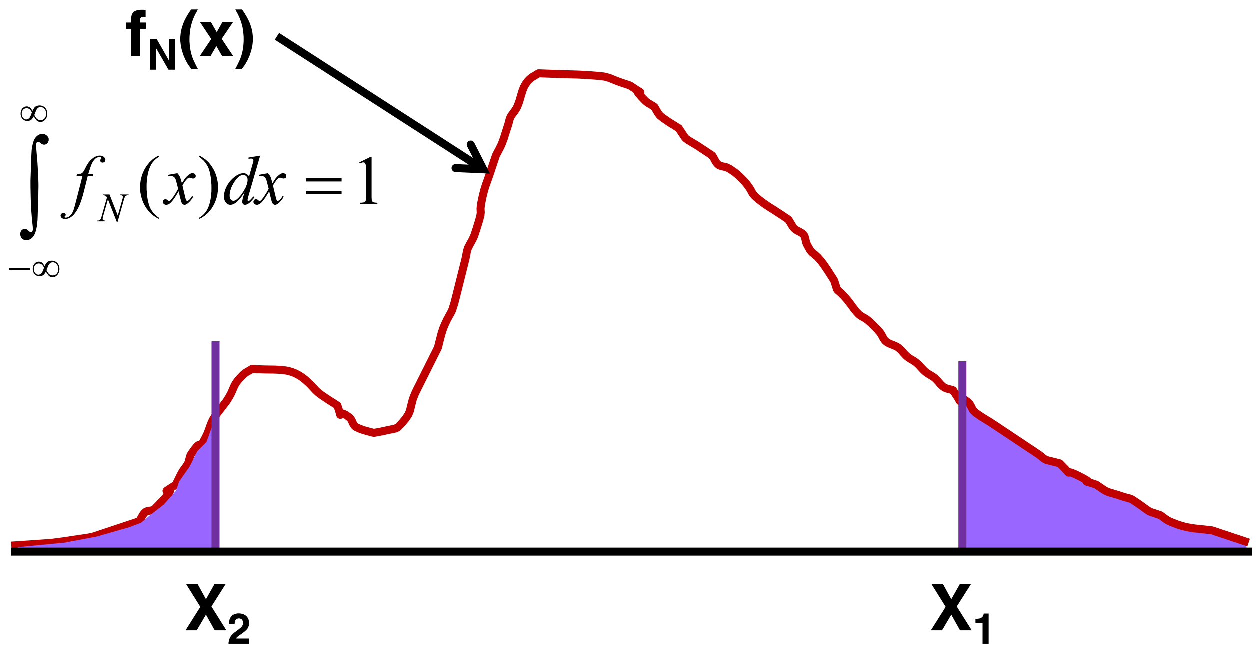
- The relative expression levels of a gene are measured in a microarray experiment testing **Drug A**

- The relative expression of Drug A versus control is reported (X_1, X_2, …, X_m)
  - If X_i = 1.4, the gene is 1.4x more expressed in drug vs control

    X = {1.2, 1.8, 1.0, 1.7, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.9 }

- Question: Does Drug A effect the expression of these genes?

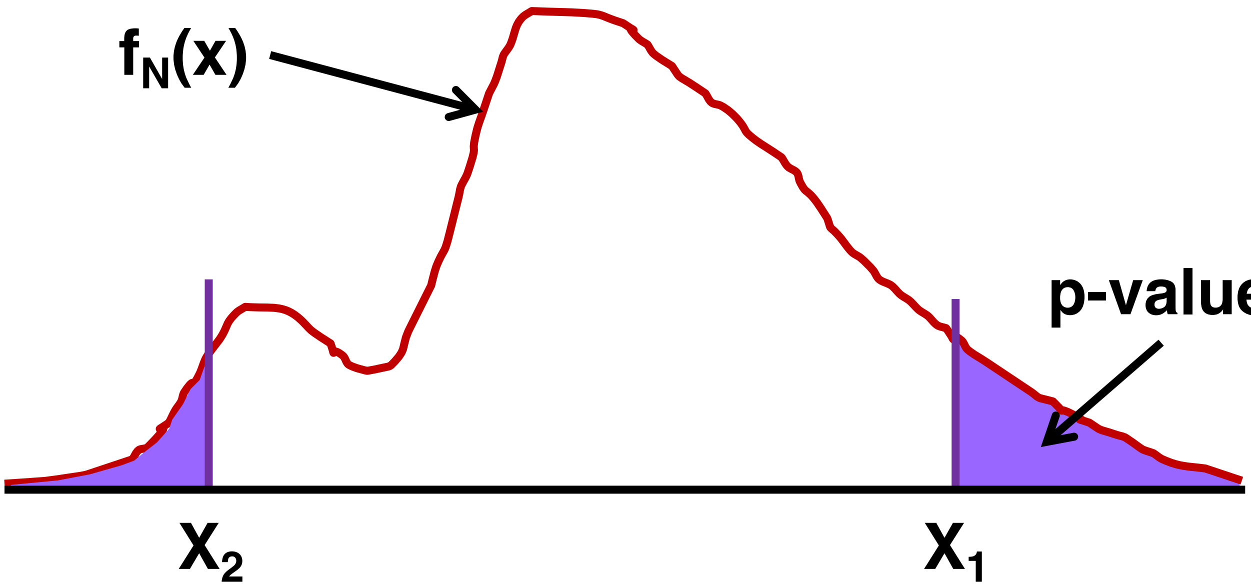
P-Value



$$p(x \le X_2) = \int_{-\infty}^{X_2} f_N(x)dx$$

$$p(x \ge X_1) = \int_{X_1}^{\infty} f_N(x)dx$$

P-Value

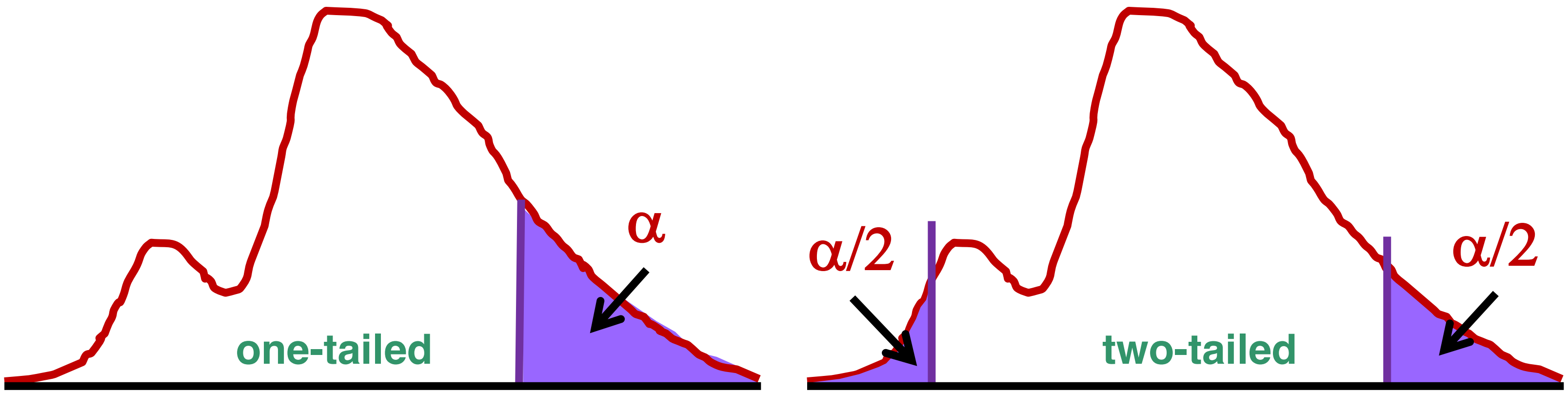


Given X_i and some distribution $f_N(X)$, the p-value is:

- **$P(x>X_i)$ (right-tailed)**
- **$P(x<X_i)$ (left-tailed)**
- **$\min\{P(x>X_i), P(x<X_i)\}$ (double-tailed)**

In hypothesis testing, $f_N(X)$ is called the null distribution

Significance Level



Alternatively, we can choose a p-value threshold, α

X1 (X2) falls into the shaded region(s) if

$P(x \geq X1) \leq \alpha$ or $P(x \leq X2) \leq \alpha$ one-tailed

or

$P(x \geq X1 \text{ or } x \leq X2) \leq \alpha/2$ two-tailed

α is called the significance level

Hypothesis Testing

- Declare Null Hypothesis H_0 and Alternative hypothesis H_1
- Decide on significance level, α
- Select a **test statistic** (and associated null distribution)
- Calculate p-value based on data
- Reject H_0 if p-value $< \alpha$.

Back to Example

- The relative expression levels of a gene are measured by microarray testing **Drug A**

X = {1.2, 1.8, 1.0, 1.7, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.9 }

H0: Drug does not change gene's expression
H1: Gene expression changes in drug

OK. But this is still vague. Can we be more specific?

Let's specify H0: The mean of the distribution of X=1
H1: Mean of distribution of X > 1

Are the Means Different?

We don't know the actual mean of the distribution of X

All we have are the samples X_i

Need to estimate the mean….

$$\overline{\mathrm{X}} = \frac{1}{\mathrm{n}} \sum_{i=1}^{N} X_i$$

Estimating Means for Example

X = {1.2, 1.8, 1.0, 1.7, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.9}

$$\overline{X} = 1.25$$

But what if the sampling had turned out differently?

X_{new} = {1.2, xx, 1.0, xx, 0.9, 1.7, xx, 1.4, 0.9, 1.2, 0.9}

$$\overline{X}_{new} = 1.15$$

Is the difference just an artifact of sampling?

The Null Hypothesis

- H_0: Xi are drawn from distribution with μ=1
- Test: what is P($\overline{X}$) given H_0?

We can't usually test this directly.

We need to define a test statistic whose distribution under H0 is completely known.

We will use this statistic to test the hypothesis *indirectly*

The Test Statistic

$$\text{Recall } \mathrm{E}\left[\overline{\mathrm{X}}\right] = \mu,$$

$$\text{and } \mathrm{Var}\left[\overline{\mathrm{X}}\right] = \sigma^2 / n$$

The Test Statistic under H_0

$$\text{Recall } \mathrm{E}\left[\overline{\mathrm{X}}\right] = \mu_0,$$
$$\text{and } \mathrm{Var}\left[\overline{\mathrm{X}}\right] = \sigma_0{}^2 / n$$

If X_i ~ normal
Then $\overline{X}$ ~ normal

The Test Statistic under H_0

$$\text{Recall } \mathrm{E}\left[\bar{\mathrm{X}}\right] = \mu_0,$$

$$\text{and } \mathrm{Var}\left[\bar{\mathrm{X}}\right] = \sigma_0^2 / n$$

$$\text{Then } \boxed{\frac{\bar{\mathrm{X}}-\mu_0}{\sqrt{\sigma_0^2/n}}} \sim N(0,1)$$

If X_i ~ normal
Then $\bar{X}$ ~ normal

So we just have to calculate this and compare to a standard normal… right?

The Test Statistic under H_0

$$\text{Recall } \mathrm{E}\left[\bar{\mathrm{X}}\right] = \mu_0,$$

$$\text{and } \mathrm{Var}\left[\bar{\mathrm{X}}\right] = \sigma_0^2 / n$$

$$\text{Then } \frac{\bar{\mathrm{X}} - \mu_0}{\sqrt{\sigma_0^2 / n}} \sim N(0,1)$$

If X_i ~ normal
Then $\bar{X}$ ~ normal

We don't know the true variance under H_0!
We only specified μ_0….

The Test Statistic under H_0

Recall $E\left[\bar{X}\right] = \mu_0$,

and $Var\left[\bar{X}\right] = \sigma_0^2/n$

Then $\boxed{\dfrac{\bar{X}-\mu_0}{\sqrt{\sigma_0^2/n}}} \sim N(0,1)$

Recall $(n-1)S^2/\sigma_0{}^2 \sim \chi^2_{n-1}$

Then $\boxed{\sqrt{S^2/\sigma_0{}^2}} \sim \sqrt{\chi^2/\text{dof}}$

$$\frac{\boxed{\dfrac{\bar{X}-u_0}{\sqrt{\sigma_0^2/n}}}}{\boxed{\sqrt{S^2/\sigma_0^2}}} = \frac{\bar{X}-u_0}{\sqrt{S^2/n}} \sim \frac{N(0,1)}{\sqrt{\chi^2/dof}}$$

The unknown σ_0^2 goes away. But can we get a p-value?

The Test Statistic under H_0

Recall $E\left[\bar{X}\right] = \mu_0$,

and $Var\left[\bar{X}\right] = \sigma_0^2 / n$

Then $\boxed{\dfrac{\bar{X}-\mu_0}{\sqrt{\sigma_0^2/n}}} \sim N(0,1)$

Recall $(n-1)S^2/\sigma_0{}^2 \sim \chi^2_{n-1}$

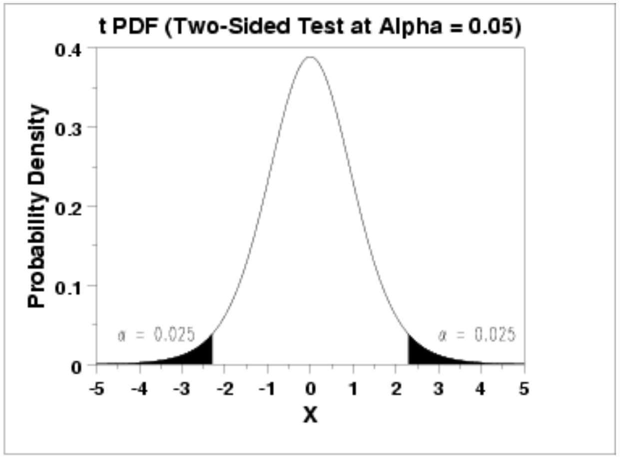
Then $\boxed{\sqrt{S^2/\sigma_0{}^2}} \sim \sqrt{\chi^2/\mathrm{dof}}$

$$\frac{\boxed{\dfrac{\bar{X}-u_0}{\sqrt{\sigma_0^2/n}}}}{\boxed{\sqrt{S^2/\sigma_0^2}}} = \frac{\bar{X}-u_0}{\sqrt{S^2/n}} \sim \frac{N(0,1)}{\sqrt{\chi^2/dof}} = t_{n-1}$$

The One Sample T-test (2-sided)

Given samples (X_1, X_2, …, X_m)
H0: Sample distribution $u = u_0$
H1: Sample distribution $u <> u_0$

$$T = \frac{\bar{X} - u_0}{\sqrt{S^2 / n}} \sim t_{n-1}$$

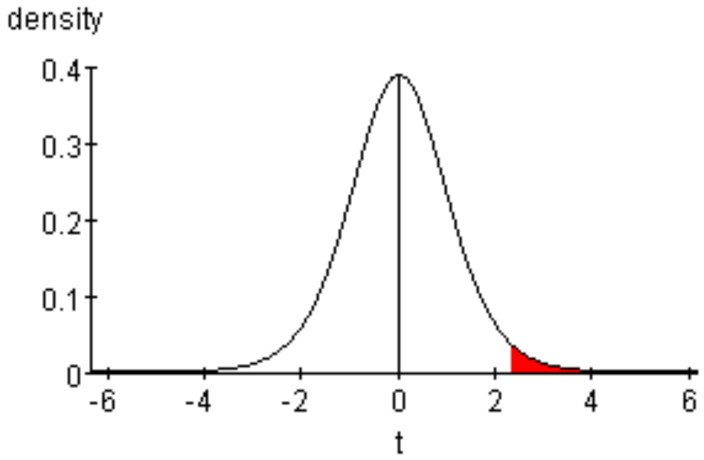


If $P(t>T)<\alpha$ under t_{n-1}, reject H_0

T-test on Example Data

X = {1.2, 1.8, 1.0, 1.7, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.9}

$$T = \frac{\overline{X} - u_0}{\sqrt{S^2 / n}} = \frac{1.25 - 1}{\sqrt{0.123/11}} = 2.36$$



P=0.02

Assumptions of the T-test

- X are drawn from a normal distribution
- Samples are independent and identically distributed (iid)

Hypothesis Testing

- $\boxed{\text{Declare Null Hypothesis } H_0 \text{ and Alternative hypothesis } H_1}$
- Decide on significance level, α
- $\boxed{\text{Select a test statistic (based on associated null distribution)}}$
- Calculate p-value based on data
- Reject H_0 if p-value $< \alpha$.

H0: u=1=u_0
H1: u> 1

$$\frac{\bar{X} - u_0}{\sqrt{S^2 / n}} \sim t_{n-1}$$

These are the keys to understanding a statistical test!

Other Useful Tests

- The χ^2 Test
- Fishers Exact Test
- Hypergeometric Test
- Wilcoxon-Mann-Whitney Test
- Permutation Test
- Kolmogorov-Smirnov Test
- Sign Test