Random Variables

A random variable X is a random, quantifiable answer that answers some probabilistic question, like "how many cars will pass here in the next 10 minutes?". The RV will have some underlying theoretical PDF $f(x)$, which will collapse into a single value when observed. Consider it a quantum particle in a superposition.

$$\underbrace{f(x) = N(x;\, \mu = 5,\, \sigma = 1.5) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}_{\text{Theoretical distribution of RV}} \quad \Rightarrow \quad \underbrace{X_i = [5.345,\, 7.955,\, 3.895,\, 1.065,\, 4.701,\, 1.696, ...]}_{\text{Sample of RV}}$$

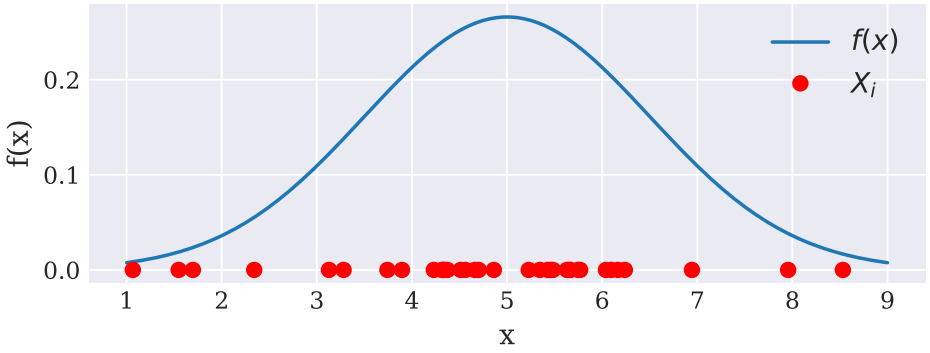

Figure 1: Random variable X, with theoretical PDF $f(x)$, and 35 observations, X_i.

Statistics

The measured sample X_i will have statistics such as mean and standard deviation, mirroring the "actual" values of the theoretical PDF. They will approach the theoretical values as the sample size increases ($N \gg 1$). The most used statistics are the mean $\overline{X}$ and variance S^2.

Properties of PDF - $f(x)$			**Statistics of RV -** X_i	
Mean	μ	$\leftrightarrow$	$\overline{X}$	Sample Mean
Variance	V	$\leftrightarrow$	S^2	Sample Variance
St.Div.	σ	$\leftrightarrow$	S	Sample St.Div.

Table 1: Properties of the RVs PDF, and the corresponding statistics of a sample of the RV.

Distribution of sample statistics

Since each random sample from the RV will be different, the statistics will differ each time as well, as seen in figure 2.

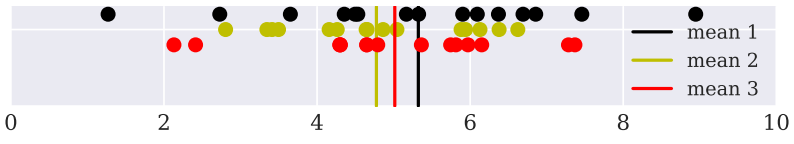

Figure 2: Three size 15 sample distributions of random variable X_i

Imagine picking an infitine amount of such random samples from the RV, each of size n. The statistics themselves will now have a distributions, with it's own mean, variance, etc.

Sample mean and Central Limit Theorem

The sample mean of the RV, independently of what sort of distribution the random samples are from, follow

$$E(\overline{X}) = \mu_{\overline{X}} = \mu \qquad V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n \tag{1}$$

Sample mean of normally distributed RV

When the RV has a **normal distribution**, the sample mean $\overline{X}$ is **itself normally distributed**, with mean and variance as in 1:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

A problem here is that σ is often unknown, as it belongs to the theoretical PDF, not the sample. The **t distribution** solves the problem. The following random variable has a t-distribution when $\overline{X}$ is the sample mean of a normal RV:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Sample mean of any RV - The Central Limit Theorem (CLT)

When the RV has **any distribution**, the sample mean $\overline{X}$ will **approach a normal distribution** for large n. This is called the central limit theorem. We can then use both the Z and T distributions as above.

Point Estimators

Estimators $\hat{\theta}$ are attempts at reconstructing a theoretical parameter (left side of 1 - μ, σ,...) from the sample statistics (right side of 1 - $\overline{X}$, S,...). The most obvious estimators are simply the mirroring statistics, but there are more advanced ones. Some definitions:

- **Mean Square Error:** $\text{MSE} = E\Big[(\hat{\theta} - \theta)^2\Big] = V(\hat{\theta}) + E(\hat{\theta}) - \theta = \text{variance of estimator} + \text{bias}^2$
- **Unbiased Estimator:** Estimator which variance is zero $V(\hat{\theta}) = 0$.
- **Minimum Variance Unbiased Estimator:** Estimator with lowest variance, which has no bias $E(\hat{\theta}) = \theta$.

Property	**MVUE of norm. dist.**	Other Estimators
Mean μ	$\overline{X}$	$\tilde{X}$, $\overline{X}_{tr(10)}$
Variance σ^2	S^2	
St.Div. σ	–	

Table 2: Most common estimators of common properties

Bootstrap point estimation

Moment estimators

Maximum likelihood estimators

Confidence Intervals

We now know what sort of distributions our estimators have. We can then establish some interval around the estimators mean where we are $100(1-\alpha)\%$ sure that the actual value lies.

Deriving Confidence Intervals

1. Find some RV which involves only your estimator (and known values), and has a known probability distribution. Ex, for the estimator $\hat{\mu} = \overline{X}$ we can use the RV

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

   which has a normal distribution, as long as σ is known, or

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

   which has a T distribution. if σ is not known.
2. Use the known distributions' percentiles to establish an interval. Ex, for the normal case:

$$P\left(z_{1-\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 100(1-\alpha)\%$$

3. Solve the inequality for the value to be estimated, in this case μ:

$$\overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad \mu \in \left[\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Useful RVs for finding for finding CIs

Distribution	**Random Variable**	Estimator	Comments
Normal Dist.	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$\hat{\mu} = \overline{X}$	Requires that σ is known.
T Dist.	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$\hat{\mu} = \overline{X}$	$n-1$ df. Approaches Z as $N \gg 1$.
Chi-squared Dist.	$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$	$\hat{\sigma^2} = S^2$	$n-1$ df. Take sqrt to get CI for σ.

Table 3: RV distributions to use in confidence

Parametric Bootstrap CI

Non-parametric Bootstrap CI