Source: R-intro.pdf (available on elearning)

8 Probability distributions

8.1 R as a set of statistical tables

One convenient use of R is to provide a comprehensive set of statistical tables. Functions are provided to evaluate the cumulative distribution function $P(X \leq x)$, the probability density function and the quantile function (given q, the smallest x such that $P(X \leq x) > q$), and to simulate from the distribution.

Distribution	**R name**	**additional arguments**
beta	`beta`	`shape1, shape2, ncp`
binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
chi-squared	`chisq`	`df, ncp`
exponential	`exp`	`rate`
F	`f`	`df1, df2, ncp`
gamma	`gamma`	`shape, scale`
geometric	`geom`	`prob`
hypergeometric	`hyper`	`m, n, k`
log-normal	`lnorm`	`meanlog, sdlog`
logistic	`logis`	`location, scale`
negative binomial	`nbinom`	`size, prob`
normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
signed rank	`signrank`	`n`
Student's t	`t`	`df, ncp`
uniform	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

Prefix the name given here by '`d`' for the density, '`p`' for the CDF, '`q`' for the quantile function and '`r`' for simulation (*r*andom deviates). The first argument is `x` for `dxxx`, `q` for `pxxx`, `p` for `qxxx` and `n` for `rxxx` (except for `rhyper`, `rsignrank` and `rwilcox`, for which it is `nn`). In not quite all cases is the non-centrality parameter `ncp` currently available: see the on-line help for details.

The `pxxx` and `qxxx` functions all have logical arguments `lower.tail` and `log.p` and the `dxxx` ones have `log`. This allows, e.g., getting the cumulative (or "integrated") *hazard* function, $H(t) = -\log(1 - F(t))$, by

```
- pxxx(t, ..., lower.tail = FALSE, log.p = TRUE)
```

or more accurate log-likelihoods (by `dxxx(..., log = TRUE)`), directly.

In addition there are functions `ptukey` and `qtukey` for the distribution of the studentized range of samples from a normal distribution, and `dmultinom` and `rmultinom` for the multinomial distribution. Further distributions are available in contributed packages, notably SuppDists.

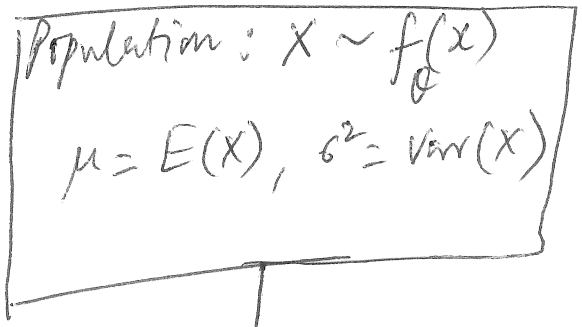
Here are some examples

```
> ## 2-tailed p-value for t distribution
> 2*pt(-2.43, df = 13)
> ## upper 1% point for an F(2, 7) distribution
> qf(0.01, 2, 7, lower.tail = FALSE)
```

See the on-line help on `RNG` for how random-number generation is done in R.

Law of large numbers and central limit theorem

Suppose the rvs $X_1, \ldots X_n$ are **independently and identically** distributed (i.i.d.) as X where $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$.



$X_1, X_2, \ldots, X_n$ (random sample).

NOTE:
- Each X_i has the same dist. as X
- The X_i are independent

- $X_1, \ldots, X_n$ represent a **random sample** of size n from the population represented by X.

- Define **sample sum**, $T = \sum_{i=1}^{n} X_i$
- $E[T] = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu$; $\text{var}[T] = \text{Var}\left[\sum_{i=1}^{n} X_i\right]$
- Define **sample average**, $\overline{X} = T/n$
- $E[\overline{X}] = E\left[\frac{T}{n}\right] = \frac{1}{n}E[T] = \frac{n\mu}{n} = \mu$.

$= \sum_{i=1}^{n} \text{var}(X_i)$ (indep.) $= \sum_{i=1}^{n} \sigma^2 = n\sigma^2$

Law of large numbers (LLN)

$$\text{var}[\overline{X}] = \text{Var}\left[\frac{T}{n}\right] = \frac{\text{var}[T]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

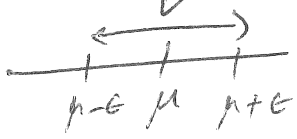
Recall:
- $E[X+Y] = E[X] + E[Y]$
- $E[aX+b] = aE(X) + b$
- $\text{var}[aX+b] = a^2 \text{Var}[X]$

Law of large #:

As $n \to \infty$, $\overline{X} \to \mu$.

- How to interpret this?

Take $\epsilon > 0$. Meaning:



- If n is large, $\overline{X} \approx \mu$.

Central limit theorem (CLT)

Version 1: If n is large, $T \approx N\left[E(T) = n\mu,\ Var(T) = n\sigma^2\right]$

($\approx$: "is approximately distributed as")

verify.

$\equiv$ If n is large, $Z = \dfrac{T - n\mu}{\sqrt{n\sigma^2}} = \dfrac{T - E(T)}{SD(T)} \approx N\left[E(Z) = 0,\ Var(Z) = 1\right)$

(Z: z-score of T)

Version: If n is large, $\overline{X} \approx N\left[E(\overline{X}) = \mu,\ Var(\overline{X}) = \sigma^2/n\right]$

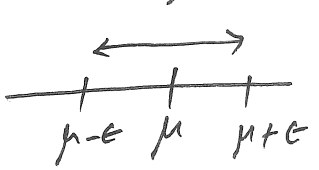
$\equiv$ If n is large, $Z = \dfrac{\overline{X} - E(\overline{X})}{SD(\overline{X})} = \dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0, 1)$.

Note: If X has normal distribution, then both T and $\overline{X}$ follow *exactly* normal distributions. (no need of CLT here).

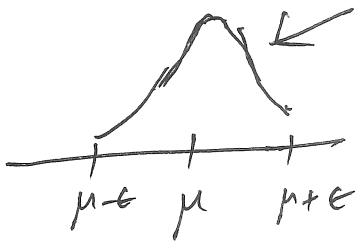
Note: The CLT holds regardless of the shape of the parent (or population dist.) However, the accuracy of the normal approx. depends on the shape of the parent dist.

Difference b/w LLN & CLT:

LLN: If n is large:



CLT: If n is large:



Population: $X \sim \text{Bernoulli}(p)$

$E(X) = p$, $\text{var}(X) = p(1-p)$

$X_1, \dots, X_n$ (random sample) → 0's and 1's

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = proportion of 1's in the sample $= \hat{p}$ ← random quantity

- $E[\hat{p}] = p$, $\text{var}[\hat{p}] = \frac{p(1-p)}{n}$ [holds for all $n \geq 1$].
- CLT: If n is large, $\hat{p} \approx N\left[E(\hat{p}) = p, \text{var}(\hat{p}) = \frac{p(1-p)}{n}\right]$

  or $Z = (\hat{p} - p) \Big/ \sqrt{p(1-p)/n} \approx N(0,1)$.