MATH411 | Fall 2018 | Chapter 3: Statistical Inference with Resampling

Dr. Yongtao Cao

TBD

Contents

3.1 Probability and Probability Distributions

Statistical inference involves drawing scientifically-based conclusions describing natural processes or observable phenomena from datasets with intrinsic **random variation**.

3.1.1 Basic Probability

Recall that the **set of all possible outcomes of a random experiment** is called a **sample space**, S. An event E is a **subset of** S.

Proposition: Law of Total Probability Let A denote an event in a sample space S, and let $B_1, B_2, \ldots, B_n$ be a disjoint partition of A. Then

$$P(A) = P(A|B_1)\, P(B_1) + P(A|B_2)\, P(B_2) + \cdots + P(A|B_n)\, P(B_n)$$

Definition A **discrete random variable** X is a function from S into the real numbers $\Re$ with a range that is **finite** or **countably infinite**. That is, $X : S \to \{x_1, x_2, \ldots, x_m\}$, or $X : S \to \{x_1, x_2, \ldots\}$.

Definition The **probability mass function** (pmf) is a function $p : \Re \to [0, 1]$ such that $p(x) = P(X = x)$, for all x in the range of X.

- Note then that $\sum_x p(x) = 1$, where the sum is over the range of X.

Definition A function X from S into the real numbers $\Re$ is a continuous random variable if there exists a non-negative function

f such that for every subset C of $\Re$, $P(X \in C) = \int_C f(x)\,dx$. In particular, for $a \leq b$, $P(a < X \leq b) = \int_a^b f(x)\,dx$.

The function f is called the **probability density function** (pdf) of X. Note that $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Definition The **cumulative distribution function** F of a random variable X is the function $F : \Re \to [0,1]$ that satisfies

$$F(x) = P(X \leq x), \quad -\infty < x < \infty$$

- F is a non-decreasing, right-continuous function with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
- In the case that X is a continuous random variable, then

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

,

and thus at every point x at which $f(x)$ is continuous,

$$F^{'}(x) = f(x)$$

by the fundamental theorem of calculus.

3.1.2 Mean and Variance

Definition: Let $X : S \to \Re$ denote a random variable and f denote its density function. The mean of X, also known as the **expected value** of X, is

$$E\left[X\right] = \mu = \int_{-\infty}^{\infty} xf\left(x\right)dx$$

and the variance is

$$Var\left[X\right] = \sigma^2 = \int_{-\infty}^{\infty} \left(x-\mu\right)^2 f\left(x\right)dx$$

Definition: The standard deviation of X is $sd\left(X\right) = \sigma_X = \sqrt{Var\left[X\right]}$.

If $X : S \to \{x_1, x_2, \ldots\}$ is a discrete random variable with $P\left(X = x_i\right) = p_i$, then

$$E\left[X\right] = \mu = \sum_i x_i \cdot p_i$$

and

$$Var\left[X\right] = \sigma^2 = \sum_i \left(x_i - \mu\right)^2 \cdot p_i$$

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

Theorem If X is a random variable with density f and g is any real-valued function, then

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

A special case of the above Theorem is: If X is a random variable and a and b are constants, then

$$E\left[a + bX\right] = a + bE\left[X\right]$$

and

$$Var\left[a + bX\right] = b^2 Var\left[X\right]$$

Definition The random variables X and Y have a joint density if there exists a non-negative function $f : \Re \times \Re \to \Re$ such that for every subset C of the plane,

$$P\left(\left(X, Y\right) \in C\right) = \iint_C f\left(x, y\right)\, dx\, dy$$

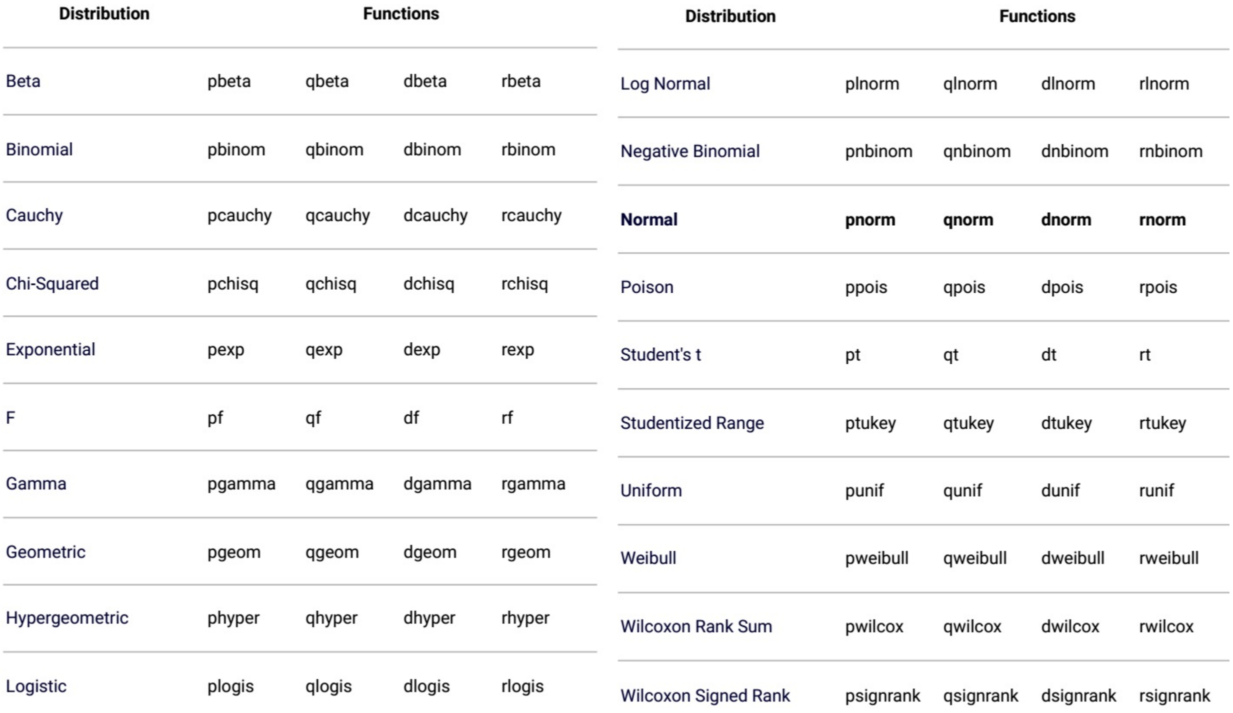
.

Definition Let X and Y be random variables. Then X and Y are independent if for any sets A and B,

$$P\left(X \in A, Y \in B\right) = P\left(X \in A\right) P\left(Y \in B\right)$$

3.1.3 Probability Distributions

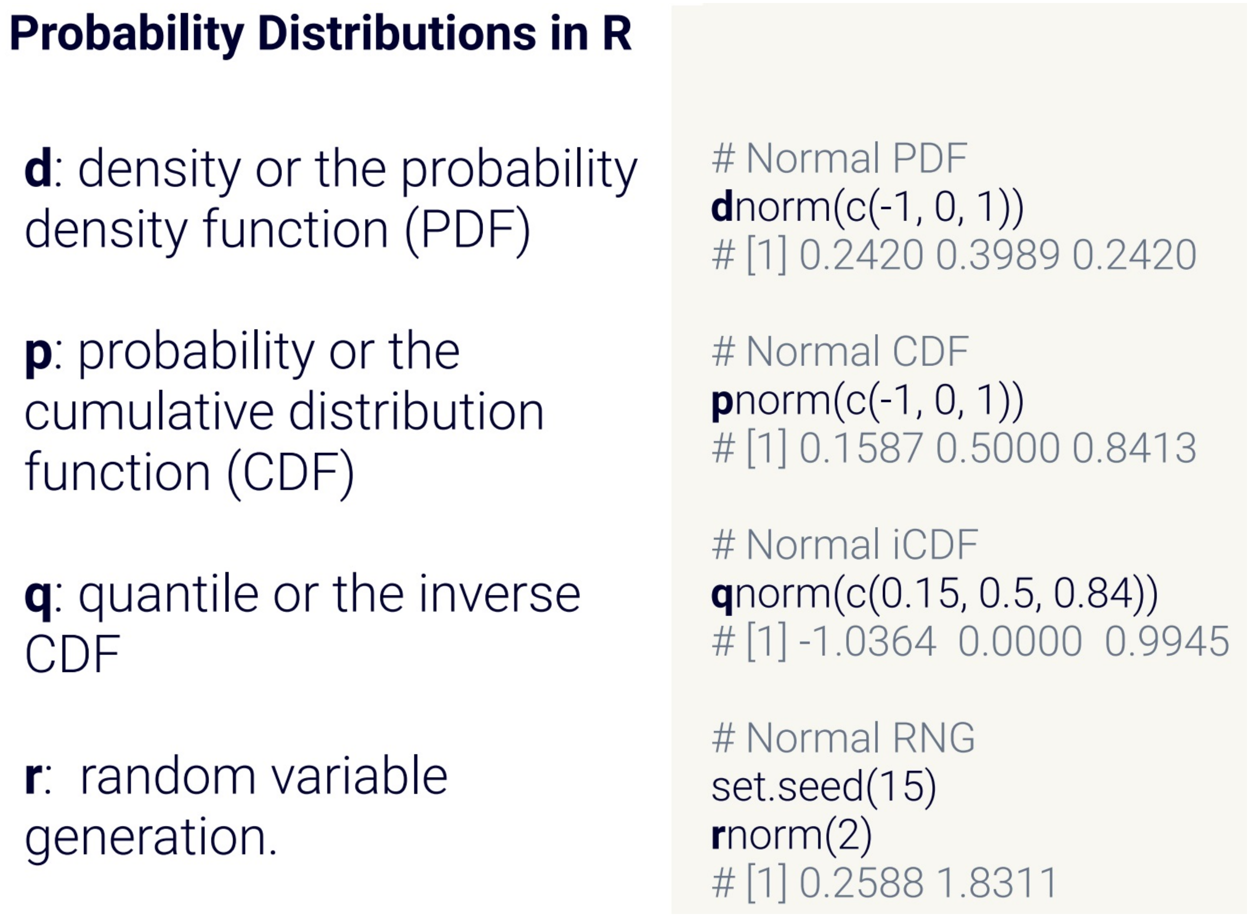
In this section, we gather results about some special probability distributions. Let's run the "distributions" APP in R.



Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Squared	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis

Distribution	Functions			
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	**pnorm**	**qnorm**	**dnorm**	**rnorm**
Poison	ppois	qpois	dpois	rpois
Student's t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank	psignrank	qsignrank	dsignrank	rsignrank

Figure 1: Some Distributions in R



Probability Distributions in R

d: density or the probability density function (PDF)

p: probability or the cumulative distribution function (CDF)

q: quantile or the inverse CDF

r: random variable generation.

```
# Normal PDF
dnorm(c(-1, 0, 1))
# [1] 0.2420 0.3989 0.2420

# Normal CDF
pnorm(c(-1, 0, 1))
# [1] 0.1587 0.5000 0.8413

# Normal iCDF
qnorm(c(0.15, 0.5, 0.84))
# [1] -1.0364  0.0000  0.9945

# Normal RNG
set.seed(15)
rnorm(2)
# [1] 0.2588 1.8311
```

Figure 2: Some Distributions in R – The Usage

Now, see R code for an example with `normal distribution`.

3.2 Permutation and Bootstrapping

Motivation Example: Does consuming beer attract mosquitoes?

A study done in Burkino Faso, Africa, about the spread of malaria investigated the connection between beer consumption and mosquito attraction. In the experiment,

- 25 volunteers consumed a liter of beer
- 18 volunteers consumed a liter of water.
- The attractiveness to mosquitoes of each volunteer was tested twice: before the beer or water and after.
- Mosquitoes were released and caught in traps as they approached the volunteers.
- The data, that is shwon in the vedio, is given in the Chapter 3 R code.

So, how do you solve this problem (Based on the skills you have learnt so far)?

We define μ_b to be the mean number of mosquitoes attracted after drinking beer and μ_w to be the mean number of mosquitoes attracted after drinking water. The hypotheses are:

$$H_0 : \mu_b = \mu_w \qquad VS \qquad H_a : \mu_b > \mu_w$$

The test statistic, assuming equal variance, is given by

$$t = \frac{\bar{x}_b - \bar{x}_w}{s_p\sqrt{1/n_b + 1/n_w}} \overset{H_0}{\sim} t_{n_b+n_w-2}$$

where $s_p = \sqrt{\frac{(n_b-1)\cdot s_b^2+(n_w-1)\cdot s_w^2}{n_b+n_w-2}}$.

Statistics Doesn't Have to be so HARD!!!

3.2.1 Permutation Test

In a permutation test, We can view our observed data as just one of many possible arrangements of the data. We can shuffle the observations around and ask "If the observations are randomly assigned treatments, what is the probability of observing our particular arrangement of the data?".

Two-sample Permutation Test Algorithm

1. Pool the $m + n$ values.
2. **repeat**

- Draw a resample of size m **without replacement**.
- Use the remaining n observations for the other sample.
- Calculate the difference in means or **another statistic** that compares samples.

until we have enough samples.

3. Calculate the P-value as the fraction of times the random statistics exceed the original statistic.

- Multiply by 2 for a two-sided test.

4. Optionally, plot a histogram of the random statistic values.

See R code for doing permutation test for the **mosquitoes** example.

Mythbusters Yawning Example: Comparing 2 Proportions

If you see someone else yawn, are you more likely to yawn? In an spisode of the show `Mythbusters`, they tested the myth that yawning is contagious.

Participants and Procideure:

- 50 adults who thought they were being considered for an appearance on the show.

- Each participant was interviewed individually by a show recruiter ("confederate") who either yawned or did not.

- Participants then sat by themselves in a large van and were asked to wait.

- While in the van, the Mythbusters watched to see if the unaware participants yawned.

Data

- 34 saw the confederate yawn (seed)
- 16 did not see the confederate yawn (control)
- 1 corresponds to yawn, 0 to no yawn

Conclusion

Finding: CONFIRMED

Really? Let's try to resolve this problem. First, let's state the hypotheses

- **Null hypothesis**: There is no difference between the seed and control groups in the proportion of people who yawned.
- **Alternative hypothesis** (directional): More people (relatively) yawned in the seed group than in the control group.

Notationally,

$$H_0 : p_{seed} = p_{control} \qquad VS \qquad H_a : p_{seed} > p_{control}$$

Based on asymptotic theory on samples, we can use

$$(\hat{p}_1 - \hat{p}_2) \sim \mathcal{N}\left(mean = (p_1 - p_2), SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

3.2.2 Bootstrapping

- Bootstrap methods are **resampling** techniques for assessing uncertainty.
- This term comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any outside help.

Bootstrapping Algorithm

1. Take a bootstrap sample – a random sample taken **with replacement** from the original data, of **the same size as the original data**.

2. Calculate the bootstrap statistic – a statistic such as `mean`, `median`, `proportion`, etc. computed on the bootstrap samples.

3. Repeat steps 1 and 2 many times to create a **bootstrap distribution** – a distribution of the bootstrap statistics.

4. **Bootstrap Percentile Confidence Intervals**: the interval between the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ percentiles of the bootstrap distribution of a statistic is a $100\,(1-\alpha)\,\%$ bootstrap percentile confidence interval for the corresponding parameter.