Tutorial Worksheet 3 - Test of Statistical Hypothesis

Problem 1. (Basics of Testing)

(a) Let Z be a random variable with probability density function $f(z) = \frac{1}{2}e^{-|z-\mu|}, z \in \mathbb{R}$ with parameter $\mu \in \mathbb{R}$. Suppose, we observe $X = \max(0, Z)$.

i. Find the constant c such that the test that "rejects when $X > c$" has size 0.05 for the null hypothesis $H_0 : \mu = 0$.

ii. Find the power of this test against the alternative hypothesis $H_1 : \mu = 2$.

(b) Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$\begin{aligned} H_0 :& \quad \theta \geq 2, \\ H_1 :& \quad \theta < 2. \end{aligned}$$

Let $Y_n = \max\{X_1, \ldots, X_n\}$, and consider a test procedure such that the critical region contains all the outcomes for which $Y_n \leq 1.5$.

i. Determine the power function of the test.

ii. Determine the size of the test.

(c) Let X be a single observation of an $\text{Exp}(\lambda)$ random variable, which has PDF

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Consider testing $H_0 : \lambda \geq \lambda_0$ versus $H_1 : \lambda < \lambda_0$.

i. Find the power function of the hypothesis test that rejects H_0 if and only if $X \geq c$.

ii. Let $0 < \alpha < 1$. Find a value of c such that the test in part (i) has size α.

iii. For what true values of λ is P_λ (type II error) $\geq 1/2$ for the test in part (i) with size α as in (ii)?

(d) Let $X_1, X_2 \overset{\text{IID}}{\sim} \text{Bin}(1, \theta)$, and consider testing $H_0 : \theta = 1/3$ versus $H_1 : \theta < 1/3$.

i. Find a test that has size 2/9 exactly. Note: It does not have to be a sensible test.

ii. Find the power function of the test from part (i), and use it to explain why this test is not a good test of these hypotheses.

(e) Suppose X is a random variable on $\{0, 1, 2, \ldots\}$ with unknown PMF $p(x)$. To test the hypothesis $H_0 : X \sim \text{Poisson}(1/2)$ against $H_1 : p(x) = 2^{-(x+1)}$ for all $x \in \{0, 1, 2, \ldots\}$, we reject H_0 if $x > 2$. The probability of type-II error for this test is

(A) $\frac{1}{4}$; (B) $1 - \frac{13}{8}e^{-1/2}$; (C) $1 - \frac{3}{2}e^{-1/2}$; (D) $\frac{7}{8}$.

Problem 2. (Likelihood ratio tests)

(a) For data $X_1, \ldots, X_n \in \mathbb{R}$ and two fixed and known values $\sigma_0^2 < \sigma_1^2$, consider the following testing problem:

$$H_0 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_0^2\right)$$
$$H_1 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_1^2\right)$$

What is the most powerful test for testing H_0 versus H_1 at level α? Letting $\chi_n^2(\alpha)$ denote the $1-\alpha$ quantile of the χ_n^2 distribution, describe explicitly both the test statistic T and the rejection region for this test.

(b) What is the distribution of this test statistic T under the alternative hypothesis H_1? Using this result, and letting F denote the CDF of the χ_n^2 distribution, provide a formula for the power of this test against H_1 in terms of $\chi_n^2(\alpha), \sigma_0^2, \sigma_1^2$, and F. Keeping σ_0^2 fixed, what happens to the power of the test as σ_1^2 increases to ∞ ?

(c) Consider two probability density functions on $[0,1]: f_0(x) = 1$ and $f_1(x) = 2x$. Among all tests of the null hypothesis $H_0 : X \sim f_0(x)$ versus the alternative $H_1 : X \sim f_1(x)$ with significance level $\alpha = 0.10$, how large can the power possibly be?

(d) Let X_1 and X_2 be a random sample from a distribution having the probability density function $f(x)$. Consider the testing of $H_0 : f(x) = f_0(x)$ against $H_1 : f(x) = f_1(x)$ based on X_1 and X_2, where

$$f_0(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

(i) For a given level, show that the critical region of the most powerful test for testing H_0 against H_1 is of the form

$$\{(x_1, x_2) : \ln x_1 + \ln x_2 > c\} \text{ for some constant } c.$$

(ii) Determine c in terms of a suitable cutoff point of a Chi-square distribution when the level is α.

(e) Let $X_1, \ldots, X_n$ be i.i.d. observation from the density,

$$f(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right), x > 0$$

where $\mu > 0$ is an unknown parameter. Consider the problem of testing the hypothesis $H_o : \mu \leq \mu_o$ against $H_1 : \mu > \mu_o$.

Show that the test with critical region $\left[\bar{X} \geq \mu_o \chi^2_{2n,1-\alpha}/2n\right]$, where $\chi^2_{2n,1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of the χ^2_{2n} distribution, has size α. Give an expression of the power in terms of the c.d.f. of the χ^2_{2n} distribution.

Problem 3. (MP tests and GLRT test)

(a) Let $X_1, \ldots, X_{10}$ be a random sample of size 10 from a population having a probability density function

$$f(x \mid \theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & \text{if } x > 1, \\ 0, & \text{otherwise} \end{cases}$$

where $\theta > 0$. For testing $H_0 : \theta = 2$ against $H_1 : \theta = 4$ at the level of significance $\alpha = 0.05$, find the most powerful test. Also find the power of this test.

(b) Let $X_1, \ldots, X_n$ be a random sample from the population having probability density function

$$f(x, \theta) = \begin{cases} \frac{2x}{\theta^2} e^{-\frac{x^2}{\theta^2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Obtain the most powerful test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(\theta_1 < \theta_0)$.

(c) Let $X_1, X_2, \ldots, X_5$ be a random sample from a $N\left(2, \sigma^2\right)$ distribution, where σ^2 is unknown. Derive the most powerful test of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 4$ against $H_1 : \sigma^2 = 1$.

(d) Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where both μ and σ^2 are unknown. Consider the problem of testing

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

Show that the generalized likelihood ratio test statistic for this problem simplifies to

$$\Lambda\left(X_1, \ldots, X_n\right) = \left(\frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 + n\bar{X}^2}\right)^{n/2}.$$

Letting $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ and $T = \sqrt{n}\bar{X}/S_X$ (the usual one-sample t-statistic for this problem), show that $\Lambda\left(X_1, \ldots, X_n\right)$ is a monotonically decreasing function of $|T|$, and hence the generalized likelihood ratio test is equivalent to the two-sided t-test which rejects for large values of $|T|$.

Problem 4. (Sign Test)

Consider data $X_1, \ldots, X_n \overset{IID}{\sim} f$ for some unknown probability density function f, and the testing problem

$$H_0 : f \text{ has median } 0$$
$$H_1 : f \text{ has median } \mu \text{ for some } \mu > 0$$

(a) Explain why the Wilcoxon signed rank statistic does not have the same sampling distribution under every $P \in H_0$. Draw a picture of the graph of a density function f with median 0, such that the Wilcoxon signed rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

(b) Consider the sign statistic S, defined as the number of values in $X_1, \ldots, X_n$ that are greater than 0. Explain why S has the same sampling distribution under every $P \in H_0$. How would you conduct a level-α test of H_0 Vs. H_1 using the test statistic S? (Describe explicitly the rejection threshold; you may assume that for $X \sim \text{Binomial}\left(n, \frac{1}{2}\right)$, there exists an integer k such that $\mathbb{P}[X \geq k]$ is exactly α.)

(c) When n is large, explain why we may reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha$ where Z_α is the upper α point of $\mathcal{N}(0, 1)$, instead of using the rejection threshold you derived in part (b).

(d) In this problem, we'll study the power of this test against the specific alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, for a fixed constant $h > 0$ (say $h = 1$ or $h = 2$) and large n. If $X \sim \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, show that

$$\mathbb{P}[X > 0] = \Phi\left(\frac{h}{\sqrt{n}}\right);$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Applying a first-order Taylor expansion of Φ around 0, show that for large n

$$\mathbb{P}[X > 0] \approx \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}.$$

(e) Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$. In this case, show that $\sqrt{\frac{4}{n}}\left(S - \frac{n}{2}\right)$ has an approximate normal distribution that does not depend on n (but depends on h) – what is the mean and variance of this normal distribution? Using this result, derive an approximate formula for the power of the sign test against the alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, in terms of Z_α, h, and the CDF Φ.

Problem 5. (Comparing Binomial proportions)

The popular search engine Google would like to understand whether visitors to a website are more likely to click on an advertisement at the top of the page than one on the side of the page. They conduct an "AB test" in which they show n visitors (group A) a version of the website with the advertisement at the top, and m visitors (group B) a version of the website with the (same) advertisement at the side. They record how many visitors in each group clicked on the advertisement.

(a) Formulate this problem as a hypothesis test. (You may assume that visitors in group A independently click on the ad with probability p_A and visitors in group B independently click on the ad with probability p_B, where both p_A and p_B are unknown probabilities in $(0,1)$.) What are the null and alternative hypotheses? Are they simple or composite?

(b) Let $\hat{p}_A$ be the fraction of visitors in group A who clicked on the ad, and similarly for $\hat{p}_B$. A reasonable intuition is to reject H_0 when $\hat{p}_A - \hat{p}_B$ is large. What is the variance of $\hat{p}_A - \hat{p}_B$? Is this the same for all data distributions in H_0 ?

(c) Describe a way to estimate the variance of $\hat{p}_A - \hat{p}_B$ using the available data, assuming H_0 is true – call this estimate $\hat{V}$. Explain heuristically why, when n and m are both large, the test statistic

$$T = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}}}$$

is approximately distributed as $\mathcal{N}(0,1)$ under any data distribution in H_0. (You may assume that when n and m are both large, the ratio of $\hat{V}$ to the true variance of $\hat{p}_A - \hat{p}_B$ that you derived in part (b) is very close to 1 with high probability.) Explain how to use this observation to perform an approximate level-α test of H_0 versus H_1.

Problem 6. (Covid-19 testing problem)

There are approximately 540 coronavirus testing locations in Abu Dhabi. At the beginning of the day, officials at each location record Y = number of specimens tested to find the first positive case and assume Y follows a geometric distribution with probability p. The probability p satisfies $0 < p < 1$ and is unknown. In this application, we might also call p the "population prevalence" of the disease. Of course, careful thought should go into defining exactly what the "population" is here.

Suppose $Y_1, Y_2, \ldots, Y_{540}$ are iid geometric (p) random variables (one for each site) observed on a given day. Epidemiologists at the Department of Health would like to test

$$H_0 : p = 0.02$$

versus

$$H_a : p < 0.02.$$

(a) Show the likelihood function of p on the basis of observing $\mathbf{y} = (y_1, y_2, \ldots, y_{540})$ is given by

$$L(p \mid \mathbf{y}) = (1-p)^{\sum_{i=1}^{540} y_i - 540} p^{540}.$$

(b) Show the uniformly most powerful (UMP) level α test of H_0 versus H_a has a rejection region of the form

$$\text{RR} = \left\{ t = \sum_{i=1}^{540} y_i \geq k^* \right\}.$$

How would you choose k^* to ensure the test is level $\alpha = 0.05$? Hint: What is the sampling distribution of $T = \sum_{i=1}^{540} Y_i$ when H_0 is true?

Problem 7. (Epidemiological Modeling of Covid-19)

SEIR models are used by epidemiologists to describe covid-19 disease severity in a population. The model consists of four different categories:

$$\begin{aligned} \text{S} &= \text{ susceptible category} \\ \text{E} &= \text{ exposed category} \\ \text{I} &= \text{ infected category} \\ \text{R} &= \text{ recovered category.} \end{aligned}$$

The four categories are mutually exclusive and exhaustive among living individuals (SEIRD models do include a fifth category for those who have died from disease). A random sample of n individuals is selected from a population (e.g., residents of Abu Dhabi city) and the category status of each individual is identified. This produces the multinomial random vector

$$\mathbf{Y} \sim \text{Multinomial}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1 \right),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}, \text{ and } \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

The random variables Y_1, Y_2, Y_3, Y_4 record the number of individuals identified in the susceptible, exposed, infected, and recovered categories, respectively. Recall that the beta distribution is a conjugate prior for the binomial. Just as the multinomial distribution can be regarded as a generalization of the binomial (to more than two categories), we need a prior distribution for $\mathbf{p}$ that is a generalization of the beta. This generalization is the Dirichlet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ distribution. Specifically, suppose $\mathbf{p}$ is best regarded as random with prior pdf

$$g(\mathbf{p}) = \begin{cases} \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} p_4^{\alpha_4-1}, & 0 < p_j < 1, \sum_{j=1}^{4} p_j = 1 \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0$, and $\alpha_4 > 0$ are known.

(a) If $Y \mid \mathbf{p} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right)$ and $\mathbf{p} \sim g(\mathbf{p})$, show the posterior distribution $g(\mathbf{p} \mid \mathbf{y})$ is Dirichlet with parameters $\alpha_j^* = y_j + \alpha_j$, for $j = 1, 2, 3, 4$. Hint: The joint distribution of $\mathbf{Y}$ and $\mathbf{p}$ satisfies

$$f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) = f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y} \mid \mathbf{p}) g(\mathbf{p})$$

and $g(\mathbf{p} \mid \mathbf{y})$ is proportional to $f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p})$.

(b) A special case of the Dirichlet distribution above arises when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, the so-called "symmetric Dirichlet distribution." This distribution would arise when one has no prior information to favor the count in one SEIR category over the other three. Do you think $g(\mathbf{p})$ with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ would be a reasonable prior model for covid-19 in Abu Dhabi city? Explain.

Problem 8. (Mean referral waiting times)

We would like to compare the population mean referral waiting times for patients in Abu Dhabi and Dubai seeking care from a gastrointestinal specialist. By "referral waiting time", I mean the time it takes to see a gastrointestinal specialist once a referral has been made by another health professional (e.g., a primary care physician, etc.). We have independent random samples of patients from the two locations. Here are the corresponding waiting times and population-level models for them:

- Abu Dhabi: $X_1, X_2, \ldots, X_m \overset{\text{IID}}{\sim}$ Exponential (θ_1)
- Dubai: $Y_1, Y_2, \ldots, Y_n \overset{\text{IID}}{\sim}$ Exponential (θ_2).

The population parameters satisfy $\theta_1 > \theta$ and $\theta_2 > \theta$ and are unknown. The goal is to test

$$H_0 : \theta_1 = \theta_2$$
$$\text{versus}$$
$$H_a : \theta_1 \neq \theta_2.$$

(a) Preparing for a LRT derivation below, carefully describe the null parameter space Θ_0 and the entire parameter space Θ. Draw a picture of what both spaces look like.

(b) Show the likelihood function is given by

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta_1^m} e^{-\sum_{i=1}^m x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^n y_j/\theta_2},$$

where $\theta = (\theta_1, \theta_2)$. This is just the likelihoods from each sample multiplied together (because the two samples are independent).

(c) Show the (restricted) MLE of θ over the null parameter space Θ_0 is

$$\hat{\theta}_0 = \begin{pmatrix} \frac{m\bar{X}+n\bar{Y}}{m+n} \\ \frac{m\bar{X}+n\bar{Y}}{m+n} \end{pmatrix}.$$

(d) Show the (unrestricted) MLE of θ over the entire parameter space Θ is

$$\hat{\theta} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}.$$

(e) Show the likelihood ratio test (LRT) statistic

$$\lambda = \frac{L\left(\hat{\theta}_0 \mid \mathbf{x}, \mathbf{y}\right)}{L(\hat{\theta} \mid \mathbf{x}, \mathbf{y})} = \frac{\bar{x}^m \bar{y}^n}{\left(\frac{m\bar{x}+n\bar{y}}{m+n}\right)^{m+n}}.$$

(f) Here are the observed data on the referral waiting times for both groups of patients:

Abu Dhabi		Dubai	
36	11	1	49
47	52	16	43
52	9	22	8
1	39	11	2
9	20	8	8
72	32	26	53
		12	24
		39	

Calculate $-2 \ln \lambda$ using the data above and implement a large-sample LRT to test H_0 versus H_a. What is your conclusion at $\alpha = 0.05$?

Problem 9. (Effect of a confounding factor)

To study the effectiveness of a drug that claims to lower blood cholesterol level, we design a simple experiment with n subjects in a control group and n (different) subjects in a treatment group. We administer the drug to the treatment group and a placebo to the control group, measure the cholesterol levels of all subjects at the end of the study, and look at whether cholesterol levels are lower in the treatment group than in the control. Let $X_1, \ldots, X_n$ be the cholesterol levels in the control group and $Y_1, \ldots, Y_n$ be those in the treatment group, and let

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

be the standard two-sample t-statistic where S_p^2 is the pooled variance. Assume, throughout this problem, that the drug in fact is *not* effective and has the exact same effect as the placebo. However, suppose there are two types of subjects, high-risk and low-risk. (Approximately half of the human population is high-risk and half is low-risk; assume that we cannot directly observe whether a person is high-risk or low-risk.) The cholesterol level for high-risk subjects is distributed as $\mathcal{N}\left(\mu_H, \sigma^2\right)$, and for low-risk subjects as $\mathcal{N}\left(\mu_L, \sigma^2\right)$.

(a) A carefully-designed study randomly selects subjects for the two groups so that each subject selected for either group is (independently) with probability 1/2 high-risk and probability 1/2 low-risk. Explain why, in this case, $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are IID from a common distribution. What are $\mathbb{E}\left[X_i\right]$ and $\operatorname{Var}\left[X_i\right]$?

(b) Explain (using the CLT and Slutsky's lemma) why, when n is large, T is approximately distributed as $\mathcal{N}(0,1)$, and hence a test that rejects for $T > Z_\alpha$ is approximately level-α for large n.

(c) A poorly-designed study fails to properly randomize the treatment and control groups, so that each subject selected for the control group is with probability p high-risk and probability $1-p$ low-risk, and each subject selected for the treatment group is with probability q high-risk and probability $1-q$ low-risk. In this case, what are $\mathbb{E}\left[X_i\right], \operatorname{Var}\left[X_i\right], \mathbb{E}\left[Y_i\right]$, and $\operatorname{Var}\left[Y_i\right]$?

(d) In the setting of part (c), show that S_p^2 converges in probability to a constant $c \in \mathbb{R}$ as $n \to \infty$, and provide a formula for c. Show that T is approximately normally distributed, and provide formulas for the mean and variance of this normal. Is the rejection probability $\mathbb{P}[T > Z_\alpha]$ necessarily close to α? Discuss briefly how this probability depends on the values μ_H, μ_L, σ^2, p, and q.

Problem 10. (Improving upon Bonferroni for independent tests)

(a) Let $P_1, \ldots, P_n$ be the p-values from n different hypothesis tests. Suppose that the tests are performed using independent sets of data, and in fact all of the null hypotheses are true, so $P_1, \ldots, P_n \stackrel{\text{IID}}{\sim} \text{Uniform}(0, 1)$. Show that for any $t \in (0, 1)$,

$$\mathbb{P}\left[\min_{i=1}^{n} P_i \leq t\right] = 1 - (1 - t)^n.$$

(b) Under the setting of part (a), if we perform all tests at significance level $1 - (1 - \alpha)^{1/n}$ (that is, we reject a null hypothesis if its p-value is less than this level), show that the probability of (falsely) rejecting any of the n null hypotheses is exactly α. Is this procedure more or less powerful than the Bonferroni procedure (of performing all tests at level α/n)?

(c) Suppose, now, that all of the p-values $P_1, \ldots, P_n$ are still independent, but not necessarily all of the null hypotheses are true. (So the p-values corresponding to the true null hypotheses are still IID and distributed as Uniform $(0, 1)$.) If we perform all tests at significance level $1 - (1 - \alpha)^{1/n}$, does this procedure control the familywise error rate (FWER) at level α? (Explain why, or show a counterexample.)

Some useful distributional functions[1]:

1. If $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then
$$Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1)$$
2. $Y \sim \mathcal{N}(0,1) \Longrightarrow Y^2 \sim \chi^2(1)$
3. $Y \sim \mathcal{N}\left(\mu, \sigma^2\right) \Longrightarrow aY + b \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right)$
4. $Y \sim \mathcal{U}(0,1) \Longrightarrow -\ln Y \sim \text{exponential}(1)$

   Generalization: $Y \sim \mathcal{U}(0,1) \Longrightarrow -\beta \ln Y \sim \text{exponential}(\beta)$

   Related: $Y \sim \text{beta}(\alpha, 1) \Longrightarrow -\ln Y \sim \text{exponential}(1/\alpha)$

   Related: $Y \sim \text{beta}(1, \beta) \Longrightarrow -\ln(1-Y) \sim \text{exponential}(1/\beta)$
5. $Y \sim \text{exponential}(\alpha) \Longrightarrow Y^{1/m} \sim \text{Weibull}(m, \alpha)$

   Related: $Y \sim \text{Weibull}(m, \alpha) \Longrightarrow Y^m \sim \text{exponential}(\alpha)$
6. $Y \sim \mathcal{N}\left(\mu, \sigma^2\right) \Longrightarrow e^Y \sim \text{lognormal}\left(\mu, \sigma^2\right)$ or equivalently if $U \sim \text{lognormal}\left(\mu, \sigma^2\right) \Longrightarrow \ln U \sim \mathcal{N}\left(\mu, \sigma^2\right)$
7. $Y \sim \text{beta}(\alpha, \beta) \Longrightarrow 1 - Y \sim \text{beta}(\beta, \alpha)$
8. $Y \sim \mathcal{U}(-\pi/2, \pi/2) \Longrightarrow \tan Y \sim \text{Cauchy}$
9. $Y \sim \text{gamma}(\alpha, \beta) \Longrightarrow cY \sim \text{gamma}(\alpha, \beta c)$, where $c > 0$

   Special case: $2Y/\beta \sim \chi^2(2\alpha)$
10. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli } (p) \Longrightarrow \sum Y_i \sim b(n, p)$
11. $Y_i \sim \text{gamma}\left(\alpha_i, \beta\right), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum Y_i \sim \text{gamma}\left(\sum \alpha_i, \beta\right)$$

    Special case: $\alpha_i = 1$, for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{exponential}(\beta) \Longrightarrow \sum Y_i \sim \text{gamma}(n, \beta)$

    Special case: $\alpha_i = \nu_i/2, \beta = 2$. Then $Y_i \sim \chi^2\left(\nu_i\right), i = 1, 2, \ldots, n$ (mutually independent) $\Longrightarrow \sum Y_i \sim \chi^2\left(\sum \nu_i\right)$

    Combination: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{exponential}(\beta)$, then
$$\frac{2\sum Y_i}{\beta} \sim \chi^2(2n)$$
12. $Y_i \sim \text{Poisson}\left(\lambda_i\right), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum Y_i \sim \text{Poisson}\left(\sum \lambda_i\right)$$
13. $Y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, n$ (mutually independent)
$$\Longrightarrow \sum a_i Y_i \sim \mathcal{N}\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$$

    Special case: $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
$$\Longrightarrow \sum a_i Y_i \sim \mathcal{N}\left(\mu \sum a_i, \sigma^2 \sum a_i^2\right)$$

[1] See also Univariate Distribution Relationships

<u>Special case of iid result</u>: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, then $\bar{Y} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$

<u>Special case of iid result</u>: If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, then $\sum Y_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$

14. If $Y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, n$ (mutually independent), then

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i} \sim \mathcal{N}(0,1),$$

for $i = 1, 2, \ldots, n$. Therefore, $U = \sum Z_i^2 \sim \chi^2(n)$ because $Z_1^2, Z_2^2, \ldots, Z_n^2$ are iid $\chi^2(1)$

15. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ geometric$(p) \Longrightarrow U = \sum Y_i \sim \text{nib}(n, p)$

16. $Y_1, Y_2 \overset{iid}{\sim} \mathcal{N}(0,1) \Longrightarrow U = Y_1/Y_2 \sim$ Cauchy

17. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim}$ exponential$(\beta) \Longrightarrow Y_{(1)} \sim$ exponential (β/n)

18. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Weibull}(m, \alpha) \Longrightarrow Y_{(1)} \sim \text{Weibull}(m, \alpha/n)$

19. If $Z \sim \mathcal{N}(0,1), W \sim \chi^2(\nu)$, and $Z \perp\!\!\!\perp W$, then

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu)$$

20. $Y_1, Y_2, \ldots, Y_n \sim$ iid $\mathcal{N}\left(\mu, \sigma^2\right)$

$$\Longrightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

21. If $W_1 \sim \chi^2\left(\nu_1\right), W_2 \sim \chi^2\left(\nu_2\right)$, and $W_1 \perp\!\!\!\perp W_2$, then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F\left(\nu_1, \nu_2\right)$$

22. If $F \sim F\left(\nu_1, \nu_2\right)$, then $1/F \sim F\left(\nu_2, \nu_1\right)$

23. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$

24. If $W \sim F\left(\nu_1, \nu_2\right)$, then

$$\frac{\left(\nu_1/\nu_2\right) W}{1 + \left(\nu_1/\nu_2\right) W} \sim \text{beta}\left(\nu_1/2, \nu_2/2\right)$$